HELSINKI UNIVERSITY OF TECHNOLOGY
Department of Electrical and Communications Engineering
Communications Laboratory

Thomas Casey

Base Station Controlled Load Balancing with Handovers in Mobile WiMAX

Master's Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Technology.

January 10th, 2008

Supervisor: Professor Riku Jäntti
Instructor: Nenad Veselinovic

HELSINKI UNIVERSITY OF TECHNOLOGY Abstract of Master's Thesis

Author: Thomas Casey
Name of the Thesis:
Base Station Controlled Load Balancing with Handovers in Mobile WiMAX
Date: January 10th, 2008 **Pages:** 109
Department: Department of Electrical and Communications Engineering
Major Study: Telecommunications
Supervisor: Professor Riku Jäntti, D.Sc.
Instructor: Nenad Veselinovic, D.Sc.

The purpose of this thesis is to examine how load balancing with Base Station initiated directed handovers could be conducted in Mobile WiMAX and the potential it has to enhance Resource Utilization and QoS system wide. An additional goal of the thesis is also to conduct preliminary research on how guard bands for rescue handovers could be used in Mobile WiMAX, how this would affect load balancing and how these two approaches could be combined.

The thesis includes a background study on the key system aspects of the IEEE 802.16e radio interface technology and WiMAX Forum Access Network Architecture in terms of load balancing and handovers and a literary review on load balancing, and system wide handover and traffic prioritization.

Based on the gained knowledge a basic Resource Utilization based load balancing algorithm tailored for Mobile WiMAX is designed. Few preliminary enhancement proposals are also made in terms of e.g. automatic tuning of the triggering threshold, multiple threshold based triggering and Resource Reservation based triggering where load balancing can be triggered in relations to the reserved guard for rescue handovers and higher priority traffic.

Finally preliminary evaluation of the basic algorithm in a static environment is conducted. Although the simulations are not extensive, beneficial information is obtained of the basic parameters of the algorithm and of the overall performance of the algorithm. Even though the basic algorithm performed well in the simulated environment, a clear need was recognized for the enhancements introduced earlier.

All in all this thesis should form a very good basis for the further development and evaluation of handover based load balancing in Mobile WiMAX. Based on the study it was concluded that load balancing with directed handovers can be a very efficient way to release resources in most cases but the use of rescue handover guard bands should still be considered.

Keywords: Load balancing, Mobile WiMAX, IEEE 802.16e, Handover prioritization, Traffic prioritization

TEKNILLINEN KORKEAKOULU Abstrakti diplomityöstä

Tekijä: Thomas Casey
Työn nimi:
Kuorman tasaus tukiaseman aloitteesta yhteysvastuun vaihdoilla mobiili WiMAX:ssa
Päivämäärä: 10.01.2008 **Sivumäärä:** 109
Osasto: Sähkö- ja tietoliikennetekniikan osasto
Pääaine: Tietoliikennetekniikka
Työn valvoja: Professori Riku Jäntti, TkT
Työn ohjaaja: Nenad Veselinovic, TkT

Tämän diplomityön päätavoitteena on tutkia, kuinka kuorman tasaus voidaan suorittaa tukiaseman aloitteesta yhteysvastuun vaihdoilla mobiili WiMAX:ssa ja selvittää menetelmän potentiaalia edistää resurssien käyttöä sekä palvelun laatua koko systeemissä. Tavoitteena on myös tutkia alustavasti sitä, miten turvakaistoja voitaisiin varata ns. pelastavalle yhteysvastuun vaihdolle mobiili WiMAX:ssa, kuinka tämä vaikuttaisi kuorman tasaukseen ja kuinka nämä lähestymistavat voitaisiin yhdistää.

Diplomityö sisältää koosteen IEEE 802.16e radiorajapintateknologian ja WiMAX Forum liityntäverkkoarkkitehtuurin tärkeimmistä elementeistä kuorman tasauksen ja yhteysvastuun vaihdon suhteen sekä kirjallisuuskatsauksen kuorman tasauksesta, sekä pelastavan yhteysvastuun vaihdon ja liikenteen priorisoinnista.

Näiden perusteella suunniteltiin mobiili WiMAX:lle räätälöity resurssien käyttöön perustuva peruskuormantasausalgoritmi. Tämän lisäksi tehtiin muutama alustava ehdotus perusalgoritmia edistävistä menetelmistä. Näihin kuuluivat esimerkiksi kuorman tasauksen laukaisuun tarkoitetun kynnyksen automaattinen säätäminen, useiden kynnysten käyttäminen sekä resurssien varaukseen perustuva laukaisu, missä kuorman tasaus voidaan laukaista turvakaistojen suhteen.

Lopuksi perusalgoritmi evaluoitiin staattisessa ympäristössä. Vaikka suoritetut simulaatiot eivät olleet laajamittaisia, perusalgoritmin parametreista ja yleisestä suorituskyvystä saatiin hyödyllistä informaatiota. Vaikka algoritmi suoriutui hyvin simuloidussa ympäristössä, aikaisemmin suunnitelluille edistäville menetelmille todettiin yleisesti ottaen selvä tarve.

Tämän diplomityön pitäisi luoda hyvä pohja yhteysvastuun vaihtoon perustuvan kuorman tasauksen edelleen kehittämiselle ja evaluoinnille mobiili WiMAX:ssa. Tutkimuksen perusteella päädyttiin siihen johtopäätökseen, että kuorman tasaus yhteysvastuun vaihdolla voi olla todella tehokas tapa vapauttaa resursseja suurimmassa osassa ympäristöistä, mutta että turvakaistojen käyttöä tulisi silti harkita.

Avainsanat: Kuormantasaus, mobiili WiMAX, IEEE 802.16e, Yhteysvastuun vaihtojen priorisointi, Liikenteen priorisointi

Preface

This thesis was carried out for the EB Corporation as a joint project with the Communications Laboratory of the Helsinki University of Technology and I'd like to thank everybody who have been involved, one way or another, in this process.

This has been a long but a very educating journey for me that has included times where tremendous perseverance and patience were required but also times that illuminated pure enjoyment from my part. I feel like I have learned great deal of new things during the course of this thesis and am eager to continue exploring the intriguing world of Wireless Communications and Radio Resource Management.

I would like to express my gratitude for my supervisor Riku Jäntti for sharing his expertise with me and wish to thank my instructor Nenad Veselinovic for his guidance. I would also like thank all my colleagues from the Communications Laboratory and EB for providing me a prosperous and well spirited working environment.

I would especially like to thank all my Indian friends from EB, Raja, Girish and Jeevaka, for forcing me to have coffee breaks once and a while, for the wonderful chats we had and for keeping my mind refreshed when I needed it the most. Special thanks to my mother, sister and my friends for all their support. Last I would like to dedicate this thesis to my lovely fiancée, Tuija, who has brought sunshine and energy to my life day after day.

Espoo, January 10th, 2008

Thomas Casey

Contents

List of Symbols and Abbreviations

$\mathbf{A}$	MS to BS association matrix
b_{max}	Maximum call blocking rate allowed
B_j^{DL}	Total throughput of all the service flows in the downlink
$c_{i,j}^{DL}$	Number of bits carried per slot in the downlink between MS j and BS i
d	Measured delay
β	Load balance index
δ	Hysteresis parameter
d_{max}	Maximum delay allowed
d_{nrt}	Maximum delay allowed for the non-real-time class
F	Radio Resource Fluctuation in a BS
F_{max}	Maximum Radio Resource Fluctuation
F_{sys}	Average Radio Resource Fluctuation in the system
G	Guard band
$G_{nrt,ho}$	Guard band for non-real-time rescue handovers
$G_{rt,ho}$	Guard band for real-time rescue handovers
$G_{rt,new}$	Guard band for new real-time flows
h	Measured handover rate
h_{max}	Maximum handover rate allowed
h_{nrt}	Maximum handover rate allowed for the non-real-time class

h_{rt}	Maximum handover rate allowed for the real-time class
h_{sen}	Maximum handover rate allowed for the most delay and handover sensitive class
λ_{rel}	Load balancing scheme slot release rate
λ_{res}	Average arrival rate of new slot reservations
L	Average load
L_{max}	Maximum average load in the system to conduct load balancing
L_{min}	Minimum average load in the system to conduct load balancing
new_avg_U	New resulting Resource Utilization after a load balancing handover
R	Reserved resources
r	Measured packet dropping rate
r_{max}	Maximum packet dropping rate allowed
r_{nrt}	Maximum packet dropping rate allowed for the non-real-time class
S_{fps}	Main frame rate
T	Triggering threshold for load balancing
T_R	Resource Reservation based triggering threshold
$T_{R,nrt}$	Resource Reservation based triggering threshold for the non-real-time class
$T_{R,rt}$	Resource Reservation based triggering threshold for the real-time class
T_U	Resource Utilization based triggering threshold
$T_{U,bas,max}$	Upper bound triggering threshold for the basic algorithm
$T_{U,bas,min}$	Lower bound triggering threshold for the basic algorithm
$T_{U,max}$	Maximum for the Resource Utilization based triggering threshold
$T_{U,min}$	Minimum for the Resource Utilization based triggering threshold
$T_{U,nrt}$	Resource Utilization based triggering threshold for the non-real-time class
t_s	Average holding time of a slot

U	Resource Utilization
$U_{DL,i}$	Downlink Resource Utilization for BS i
$U_{DL,tot}$	Total number of slots in the downlink subframe
$U_{UL,i}$	Uplink Resource Utilization for BS i
U_i	Resource Utilization of BS i
3GPP	3rd Generation Partnership Project
AAA	Authentication, Authorization and Accounting
AP	Access Point
ARQ	Automatic Repeat reQuest
ASN	Access Service Network
ASN-GW	Access Service Network-Gateway
BE	Best Effort
BSC	Base Station Controller
CINR	Carrier to Interference and Noise Ratio
CPP	Cut-off Priority Policy
CQICH	Channel Quality Indication Channel
CSN	Connectivity Service Networks
CTC	Convolutional Turbo Code
DCA	Dynamic Channel Allocation
DCD	Downlink Channel Descriptor
DCF	Distributed Coordiation Function
DH	Directed Handover
DL	Downlink
DMTBR	Dynamic Multiple-Threshold Bandwidth Reservation
DR	Directed Retry
DWRR	Deficit Weighted Round Robin
ertPS	Extended Real-Time Polling Service

FBSS	Fast Base Station Switching
FCA	Fixed Channel Allocation
FDD	Frequency Division Duplexing
FIFO	First-in-First-out
FTP	File Transfer Protocol
FUSC	Fully Used Sub-Carrier
GPS	Global Positioning System
GSM	Global System for Mobile Communications
HA	Home Agent
HHO	Hard Handover
HO	Handover
HTTP	Hyper Text Transfer Protocol
IEEE	Institute of Electrical and Electronics Engineers
LB	Load Balancing
LBC	Load Balancing Cycle
MAHO	Mobile Assisted Handover
MAP	Media Access Protocol
MAP IE	Media Access Protocol Information Element
MDHO	Macro Diversity Hand Over
MRTR	Minimum Reserved Traffic Rate
MS	Mobile Station
MSS	Mobile Subscriber Station
MSTR	Maximum Sustained Traffic Rate
NAP	Network Access Provider
nrt	non-real-time
nrtPS	Non-Real-Time Polling Service
NSP	Network Service Provider

OFDMA	Orthogonal Frequency Division Multiple Access
OHHO	Optimized Hard Handover
OTcl	Object TCL
PF	Policy Function
PHY	Physical layer
PMP	Point to Multipoint
PSL	Physical Service Level
PUSC	Partially Used Sub-Carrier
QoS	Quality of Service
RNC	Radio Network Controller
RRA	Radio Resource Agent
RRC	Radio Resource Controller
RRM	Radio Resource Management
RSSI	Received Signal Strength Indicator
rt	real-time
rtPS	Real-Time Polling Service
SBS	Serving Base Station
SCR	Spare Capacity Report
SISO	Single Input Single Output
TBS	Target Base Station
TCP	Transmission Control Protocol
TDD	Time Division Duplexing
UCD	Uplink Channel Descriptor
UDP	User Datagram Protocol
UGS	Unsolicited Grant Service
UL	Uplink
UMTS	Universal Mobile Telecommunications Service

VAD	Voice Activity Detection
VoIP	Voice over Internet Protocol
WiMAX	Worldwide Interoperability for Microwave Access
WLAN	Wireless Local Area Network
WRR	Weighted Round Robin

Chapter 1

Introduction

Mobile Worldwide Interoperability for Microwave Access (WiMAX) is an IEEE broadband wireless access technology, targeted to provide a high speed wireless access for long distances, in a mobile environment. Its predecessor, fixed WiMAX, is based on the IEEE 802.16-2004 standard [IE^{3}04] and supports only limited coverage and roaming. Mobile WiMAX is based on the 802.16e-2005 amendment [IE^{3}05] to the 2004 standard and introduces true mobility to the WiMAX system. The 802.16 standard differs from the existing IEEE 802 family in the sense that it offers a very high utilization of radio resources and a good Quality of Service (QoS) framework provided by connection oriented Medium Access Control (MAC) and agile centralized scheduling.

Mobile WiMAX as a system is not equal to the 802.16e version of the standard. Only the radio link used in Mobile WiMAX is based on a subset of the features and functionalities defined in 802.16e. An organization called the WiMAX Forum is in charge of collecting this subset and integrating it to an Access Service Network (ASN) with many base stations to form the final Mobile WiMAX system profile that will be deployed by vendors.

Until now, the resource utilization and QoS within the radio link of one Base Station (BS) have received a lot of attention and have been the target of many research projects. As WiMAX evolves to include mobility, a cellular infrastructure and overlapping cells, system wide Radio Resource Management (RRM) and QoS within the ASN access network become interesting and relevant issues. The additional cost, that inefficient system wide resource utilization introduces, will form an evident problem for operators in the future. Studying these relatively new topics within the WiMAX context is therefore attractive and beneficial.

The Mobile WiMAX system profile provides a RRM framework for more efficient system wide resource utilization with the help of load balancing (LB). In Mobile WiMAX load balancing can be conducted by forcing handovers (HO) from highly loaded (”hot-spot”) Base Stations to lightly loaded ones. When a Base Station is

overloaded by traffic the QoS of many users will degrade and hence load balancing can be used to enable better QoS system wide.

For many functionalities, the 802.16 standard and the Mobile WiMAX system profile only define a framework for procedures and measurements, but leave the actual detailed implementation and algorithms to be chosen by the vendor. The same applies to load balancing and hence the main goal of this thesis is to study and evaluate how system wide load balancing could be conducted in Mobile WiMAX. The load balancing logic can reside in the Mobile Subscriber Station (MSS) or in the Base Station. In this thesis we will concentrate on, how load balancing can be controlled and initiated by the Base Station.

Mobility also introduces another important aspect of system wide QoS. Generally dropping an existing connection or lowering its QoS is considered being worse than blocking a new one. Thus when providing QoS system wide, the system has to often prioritize ongoing connections over new ones and set aside sufficient radio resources so that ongoing sessions are not dropped when the session is migrating (handed over) to another Base Station. If a Base Station is heavily loaded, if the number of MSSs in the overlapping areas is small or if the MSSs are very mobile, load balancing might not release sufficient amount of free resources for this.

Therefore another goal of this thesis is to examine how we could complement the load balancing scheme by providing such a guard band and what kind of a relationship they have. We will also study the possibility of setting different guard bands for connections using different scheduling services. Prioritizing traffic by providing different level of service is one of the fundamental features of fixed WiMAX and using such multiple thresholds could extend this concept to mobility. Such handover and traffic prioritization will also have an impact on how and when load balancing is triggered which makes their examination even more interesting.

Objectives of the thesis

To understand the goal of the thesis we will make a clear distinction between two kinds of handovers: *directed handovers* and *rescue handovers*. In a directed handover the BS tells the MSS to do a handover to a recommended Target BS (TBS). Directed handovers can be used proactively by the BS to distributed traffic load across the system[1] thus enhancing the possibility, that each individual Base Station and the system as a whole can fulfill the QoS guarantees made for the Subscriber Stations.

Rescue handovers on the other hand occur when an MSS drifts away from the Serving BS (SBS) towards a TBS. As a result of a deteriorating radio signal, the

[1]We will use the term system to refer to the cluster of Base Stations within which load balancing is conducted.

connection of the MSS has to be handed over to a TBS ("rescued" by the TBS) to guarantee sufficient signal quality.

The main goal of this thesis is to examine how load balancing with directed handovers could be done in Mobile WiMAX and what kind of potential it has to enhance resource usage and QoS system wide.

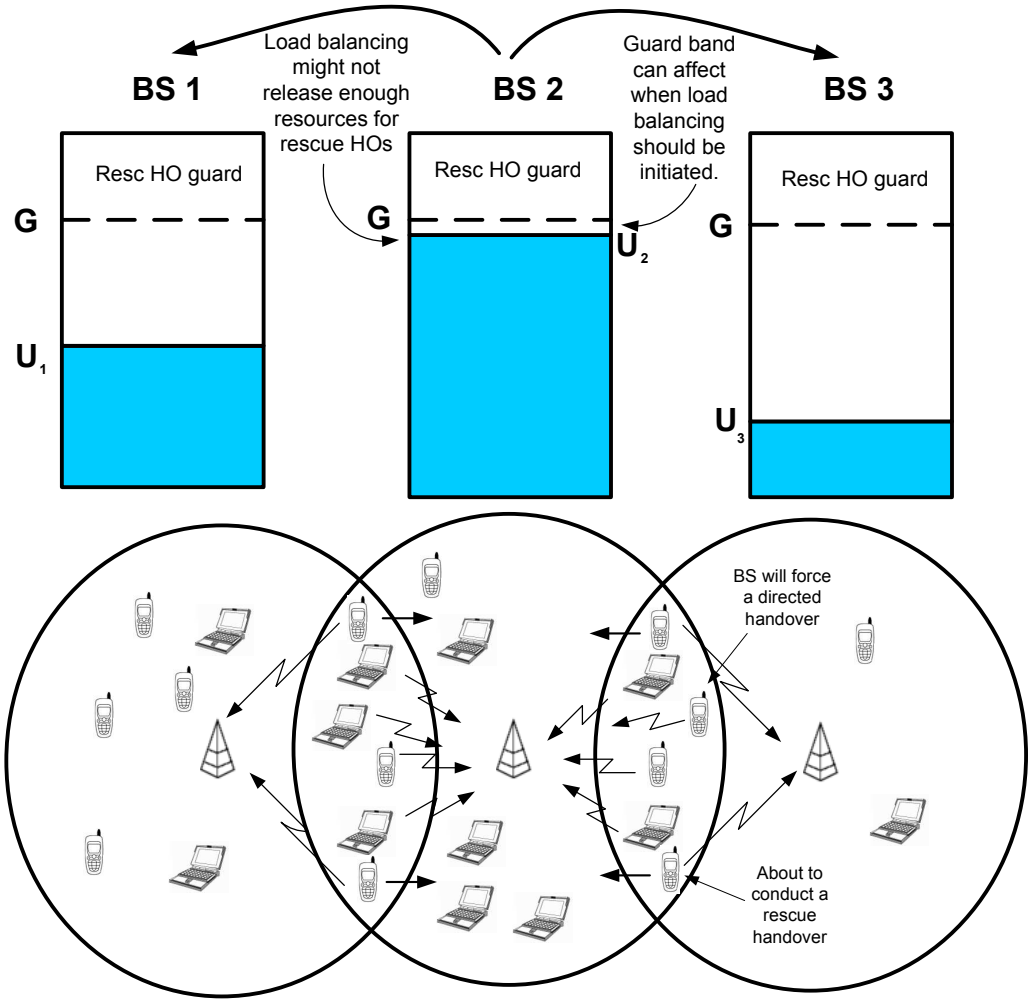

Figure 1.1: Load balancing when mobility is considered.

However, as load balancing has its limitations, it is possible that load balancing alone cannot release enough resources to eliminate rescue handover drops. Therefore an additional goal of the thesis is to conduct preliminary research on how the reservation of additional resources for rescue handovers (rescue HO prioritization) could be conducted, how it affects load balancing and how the two could be combined. Usually load balancing is triggered when a certain threshold in resource utilization

is passed but as can be seen from Figure 1.1 load balancing can be triggered in relations to the guard band reserved for rescue handovers.

Furthermore when guard bands for different kinds of traffic are introduced the load balancing problem comes even more interesting and new possibilities to enhance QoS come forth.

The objectives of this thesis include background study on the key system aspects of Mobile WiMAX that relate to load balancing and handovers and general background study on both load balancing and handover and traffic prioritization. The objective is also to design a basic load balancing algorithm (based on previous research) and to introduce a framework for enhanced load balancing algorithms featuring the effect of rescue handover prioritization and traffic type prioritization. In addition the aim is to conduct preliminary evaluation of the basic load balancing algorithm in a static environment (static MSSs). Evaluation of the basic load balancing algorithm in a mobile environment and the evaluation of the enhanced load balancing algorithms will remain outside the scope of the thesis.

To summarize our goal as a problem statement, our aim is to get an answer to the following questions:

- Overall:
  - How can BS initiated load balancing with directed handovers be used in Mobile WiMAX to enhance the utilization of radio resources?
  - How can load balancing with directed handovers be complemented with handover and traffic prioritization to guarantee system wide QoS for rescue handovers and what kind of an effect does such prioritization have on load balancing?
- More specifically:
  - What kind of a framework does Mobile WiMAX provide for load balancing with directed handovers and rescue handover prioritization?
  - What has been done previously in terms of load balancing with handovers, with rescue handover prioritization and with system wide traffic prioritization?
  - How could load balancing with directed handovers be applied to Mobile WiMAX and how could it be complemented with and connected to rescue handover prioritization and traffic prioritization? How could load balancing be triggered in relations to handover prioritization?
  - When simulated in a static environment, what kind of preliminary results can be obtained on the performance of load balancing in terms of system wide resource utilization and QoS? How should load balancing parameters be tuned?

– Could load balancing release enough resources to guarantee system wide QoS in a mobile environment where also rescue handovers are conducted or should the use of handover and traffic type prioritization be considered?

Outline of the thesis

The rest of this thesis is organized as follows. In Chapter 2 we will have an overview look of the Mobile WiMAX system in terms of its IEEE 802.16e radio link technology and the ASN access network, concentrating on the issues related to load balancing and handovers. In Chapter 3 we will do a literary review on load balancing, and system wide handover and traffic prioritization to understand the theory behind them. In Chapter 4 we will apply this theory to Mobile WiMAX, by using a basic load balancing algorithm from previous research, modifying it to be used in Mobile WiMAX and introducing possible enhancements to it. In Chapter 4 we will also conduct preliminary study of how the algorithm could be further enhanced by complementing it with handover prioritization and traffic prioritization. In Chapter 5 we will evaluate the basic load balancing algorithm in a static environment and analyze its behavior and in Chapter 6 we summarize our work, draw conclusions and take a look at possible future work.

Chapter 2

Overview of the Mobile WiMAX system

As the cell phone penetration is growing at a high rate and the demand for true mobile broadband is increasing, Mobile WiMAX offers an attractive choice to complement the existing cellular and wireless networks such as Global System for Mobile Communications (GSM), Wireless Local Area Network (WLAN) and Universal Mobile Telecommunications Service (UMTS).

Mobile WiMAX is a special system that combines both efficient radio resource utilization and a versatile QoS support on the MAC level. It therefore has the potential to serve a wide range of terminals with different needs, from static terminals that require only Best Effort (BE) scheduling services to mobile terminals that need a guaranteed bit rate for voice connections even when moving from one cell to another.

In this chapter we will take an overall glance at the radio interface technology of IEEE 802.16e and the access network used in Mobile WiMAX. We will approach both of these from the point of view of handovers, load balancing and QoS. Our aim is to get a good understanding of the framework that Mobile WiMAX offers for all of these.

2.1 Overview of the IEEE 802.16e technology

In this section we will go through the basics of the IEEE 802.16e radio interface. We will first take a look at the physical layer and the flexible Orthogonal Frequency Division Multiple Access (OFDMA) frame structure that Mobile WiMAX uses and move on to see how this enables the MAC scheduler to efficiently and flexibly use the radio resources and provide QoS within one Base Station. As load balancing extends the possibility to meet these QoS requirements, it is of interest to know the background behind them.

In this section we will also go through the handover support that IEEE 802.16e offers

to understand the functionality behind directed (BS initiated) and rescue (MS(S) initiated) handovers in Mobile WiMAX. The terminal entity in Mobile WiMAX is called a Mobile Subscriber Station (MSS), but we will from now on refer to it only as Mobile Station (MS). If not mentioned otherwise, the issues covered here will be based on [IE^{3}04], [IE^{3}05] and [WiMAX].

2.1.1 PHY layer

The physical level (PHY) of IEEE 802.16e offers a flexible frame structure and the possibility for simple and efficient frequency reuse planning. Below we will take an overview look on both.

2.1.1.1 Frame structure

The physical layer in Mobile WiMAX is based on OFDMA access technology which transmits information using a large number of closely-spaced orthogonal sub-carriers. Figure 2.1 presents the flexible OFDMA frame structure offered by IEEE 802.16e. Both Frequency (FDD) and Time Division Duplexing (TDD) are supported but TDD will most likely be the preferred choice due to its flexible support of asymmetric downlink and uplink traffic and other benefits such as less complex transceivers. In TDD the frame is divided into a downlink (DL) and uplink (UL) subframe [WiMAX]. The division is usually fixed but can also be dynamic. The subframes are divided in the time dimension to OFDMA symbols and in the frequency dimension to sub-channels.

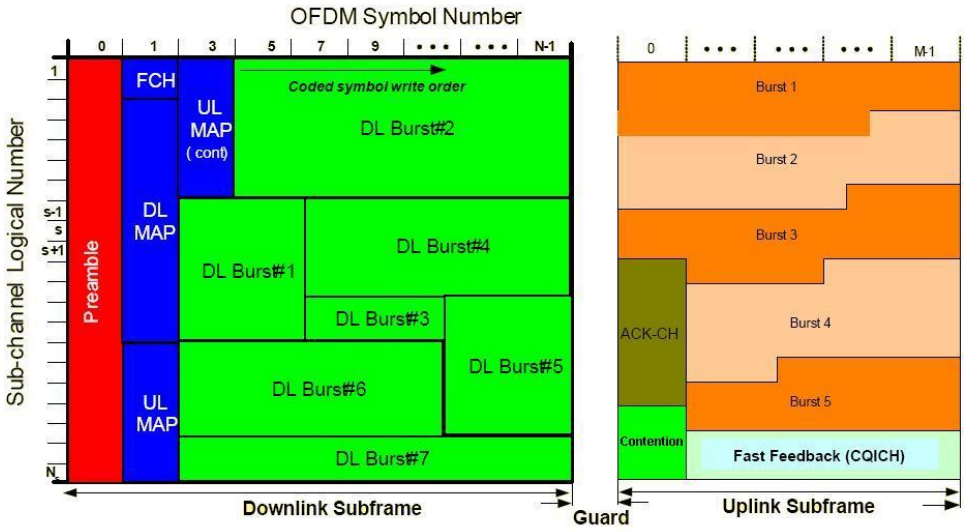

Figure 2.1: The OFDMA frame structure [WiMAX].

The flexible OFDMA frame structure that IEEE 802.16e provides, makes it possible to allocate bursts for each individual Mobile Station both in time and frequency

dimension within each individual frame enabling very high radio resource utilization.

The downlink subframe includes a preamble used for synchronization and DL- and UL-MAP (Media Access Protocol) headers that point to the places where each UL and DL burst begins so that each MS knows when and in which frequency to receive and send. The uplink subframe features dedicated channels for initial ranging procedures (used also in the network re-entry phase of a handover), contention to enable bandwidth requests to be sent for UL data transmission and feedback of the radio channel conditions (Fast Feedback Channel Quality Indication Channel (CQICH)) based on which the Modulation and Coding Scheme (MCS) can be adjusted for each MS to meet its current channel conditions (Link Adaptation).

Also an Uplink and Downlink Channel Descriptor (UCD and DCD) message describing the PHY level attributes of the BS, is broadcasted in the donwlink subframe on a periodical basis. The messages includes important frequency, power and timing information that the new MSs entering the BS needs to be able to range the BS. The size of the UCD and DCD messages is large so they are transmitted quite rarely (order of many seconds).

2.1.1.2 Frequency reuse

Mobility will introduce a cellular infrastructure with overlapping cells and handovers to WiMAX. Therefore special care must be targeted towards interference issues at the cell edges. Traditional cellular networks are notorious in terms of their need for manual radio network planning and optimization.

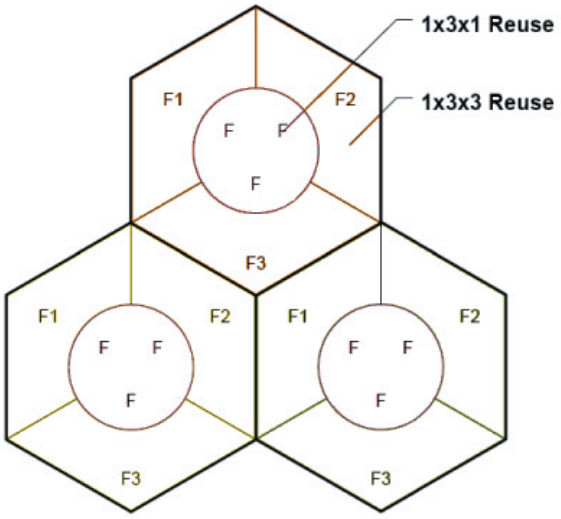

Figure 2.2: Fractional frequency reuse [Ahm06].

Mobile WiMAX aims to relief this issue by using fractional frequency reuse. The idea is to use only a part of the subchannel set (Partially Used Sub-Carrier (PUSC)) when providing a connection at the cell edge and to appropriately configure the rest of the network so that there is no need to conduct traditional frequency planning. A typical example of fractional frequency reuse 1x3x3 is shown in Figure 2.2.

Due to the flexible frame structure, the MSs located in the middle of the cell can utilize all the subchannels (Fully Used Sub-Carrier (FUSC)) as long as they are far enough from the other Base Stations. The the middle part of the cell can be therefore treated as an isolated cell, where a 1x3x1 frequency reuse scheme can be enforced, enabling even higher radio resource utilization.

The sectorization introduced by PUSC increases overlapping and enhances possibilities to conduct load balancing with directed handovers. Load balancing can be conducted, not only by doing inter-cell handovers in the overlapping areas between cells, but also by doing intra-cell handovers between the sectors.

2.1.2 MAC and QoS

The flexible MAC level scheduling and the QoS support that it enables form by far one of the most salient features in IEEE 802.16e. Following is a short overview of both.

2.1.2.1 Scheduling

In the IEEE 802.16e Point to Multipoint (PMP) operational mode, where communication is conducted only between the MS and the BS, the MAC scheduler is located in the Base Station. The standard also defines a mesh operational mode where the MSs can communicate with each other in an Ad-Hoc manner, but as this is outside the scope of this work we will not discuss it any further.

In IEEE 802.16e each MS enters the Base Station by first performing ranging and after that trying to create a service flow for its connection. After an admission control check that ensures the BS has enough resources for the service flow, the service flow is admitted to the BS, and from there on, it is the job of the scheduler to satisfy the QoS guarantees promised for the MS in the admission control check. The role of the scheduler is well described by a triangle model often used in teletraffic theory.

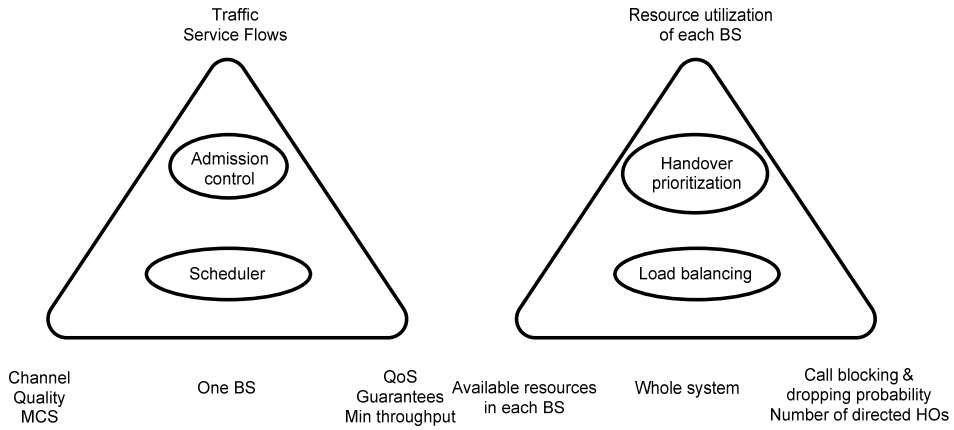

Figure 2.3: The triangle model applied to the Scheduler of one BS and to the whole system.

As can be seen, after admission control, the scheduler has to fulfill the QoS requirements of each service flow by efficiently using the time varying radio resources. This means that in case of a lack of radio resources the scheduler has to delay the transmission of lower priority service flows.

To see the resemblance between scheduling on a single BS and on system level, the same model can be applied for load balancing and handover and traffic prioritization. Both can be thought of as system level scheduling schemes, where handover prioritization serves as a kind of system level admission control trying to guarantee a certain call blocking or dropping probability[1] and load balancing as a system level scheduler trying to compensate for the effect of non-uniformly distributed traffic.

In the downlink direction packet scheduling is pretty straightforward. As the MAC scheduler is located in the BS, it knows how many packets each connection has in its queue and can therefore make decisions based on the QoS guarantees made for each MS and the channel quality information obtained from each MS. The capacity of the basic MAC resource unit, slot[2], can change since the Modulation and Coding Scheme will be chosen to meet the requirements of current channel conditions.

In the uplink direction scheduling is a bit more challenging. Because the MAC scheduler can not know the number of packets residing in the UL queue of the MS, the MS has to send bandwidth requests to get a permit from the scheduler to send data in the UL direction. The bandwidth request is sent either through the contention channel mentioned in subsection 2.1.1.1 or through a dedicated polling channel depending on the QoS guarantees. This way the MAC scheduler can ef-

[1] Or degradation of QoS.

[2] Slot is the minimum frequency-time resource unit and is equal to 48 data tones (sub-carriers). A slot can be organized differently in the time-frequency dimension depending on which sub-carrier permutation (e.g. PUSC/FUSC) scheme is used.

ficiently utilize the resources in the frame and allocate data grants for each MS, where the data bursts can be sent. This sort of centralized coordination utilizes the radio resources much more efficiently than for example the 802.11 Distributed Coordiation Function (DCF) mode where data collision occur frequently.

2.1.2.2 QoS framework

With the help of efficient MAC level scheduling, the IEEE 802.16e is able offer connection oriented QoS in addition to traditional BE scheduling services. The standard defines five different scheduling services that are summarized in Table 2.1.

Table 2.1: The scheduling services.

Scheduling Class	Typical applications	QoS parameters
UGS	VoIP	Maximum Sustained Traffic Rate Maximum Latency Tolerated Jitter
rtPS	Streaming Video & Audio	Minimum Reserved Traffic Rate Maximum Sustained Traffic Rate Maximum Latency Tolerated Jitter Traffic Priority
ertPS	VoIP with VAD	Minimum Reserved Traffic Rate Maximum Sustained Traffic Rate Maximum Latency Tolerated Jitter Traffic Priority
nrtPS	critical FTP and HTTP	Minimum Reserved Traffic Rate Maximum Sustained Traffic Rate Traffic Priority
BE	HTTP	Maximum Sustained Traffic Rate Traffic Priority

The Unsolicited Grant Service (UGS) is meant for service flows that generate fixed-size data packets on a fixed periodic interval such as Voice over Internet Protocol (VoIP). UGS gives unsolicited data grants, does not even use the bandwidth request mechanism and is therefore ideal to services that are very delay and jitter sensitive.

Real-Time Polling Service (rtPS) scheduling service is intended for service flows that generate variable size data packets on a fixed periodic interval such as streaming video and audio. A MS with a service flow using rtPS will be polled on a periodic basis so it can specify the amount of data the service flow has to send.

The Extended Real-Time Polling Service (ertPS) scheduling service combines the strengths of the UGS and rtPS scheduling classes. Like in UGS it gives data grants in an unsolicited way but the Data Grant sizes can be dynamically changed (e.g. to zero) in manner similar to the rtPS class. It is mainly aimed for VoIP with Voice Activity Detection (VoIP with VAD).

The Non-Real-Time Polling Service (nrtPS) scheduling service is targeted for service flows that are critical but not time sensitive such as File Transfer Protocol (FTP) transfers. In other words nrtPS guarantees a minimum throughput but no guarantees in terms of delay or jitter are made.

The Best Effort (BE) scheduling service is designed for service flows that carry best effort traffic such as every day web browsing. The different parameters of these QoS services will play an important role when evaluating the efficiency of load balancing and when providing system wide QoS in the case of a rescue handover. What is especially important in our case is the minimum guaranteed throughput[3].

2.1.3 Handovers

Handovers are the essential element of system wide resource utilization and QoS and it is therefore of interest to us to know the background behind them. On the other hand they are a tool for us to use resources more efficiently system wide (directed handovers) but on the other hand a burden that we have to prepare for (rescue handovers).

So how are handovers actually conducted in IEEE 802.16e? The question is not a straightforward one to answer, because handovers in IEEE 802.16e can be done in many ways. All handover types, from an uncontrolled handover to a network controlled and optimized handover, are supported.

The handover in IEEE 802.16e can be roughly divided into four phases:

1. Cell re-selection,
2. Handover decision and initiation (followed by resource reservation and admission control by the Target BS),
3. Synchronization to the Target BS downlink (interruption in the connection) and
4. Network re-entry (including ranging)(interruption in the connection).

[3]Maximum Sustained Traffic Rate (MSTR) in the case of UGS and ertPS and Minimum Reserved Traffic Rate (MRTR) in the case of rtPS and nrtPS. The bandwidth from MRTR to MSTR is not guaranteed for rtPS and nrtPS but allocated if enough resources are available.

The two typical reasons to initiate a handover that are of interest to us[4] are:

- MS initiated *rescue handovers* where the connection of the MS has to be handed over to a TBS as a result of a deteriorated radio signal ("rescued" by the TBS).
- BS initiated *directed* handovers where the handover is initialized as a result of an unbalanced distribution of traffic.

How will the handover phases differ for the MS initiated rescue handover and for the BS initiated directed handover? Phases 3 and 4 progress more or less in the same way but for phase 1 and 2 the procedure is different. Following is a closer look at all the phases.

2.1.3.1 Phase 1: Cell reselection

The purpose of the cell reselection phase, is for the MS to find out which Target Base Station (or many TBSs) it can handover to. The first set of potential TBSs is obtained from a network topology advertisement message (MOB_NBR-ADV).

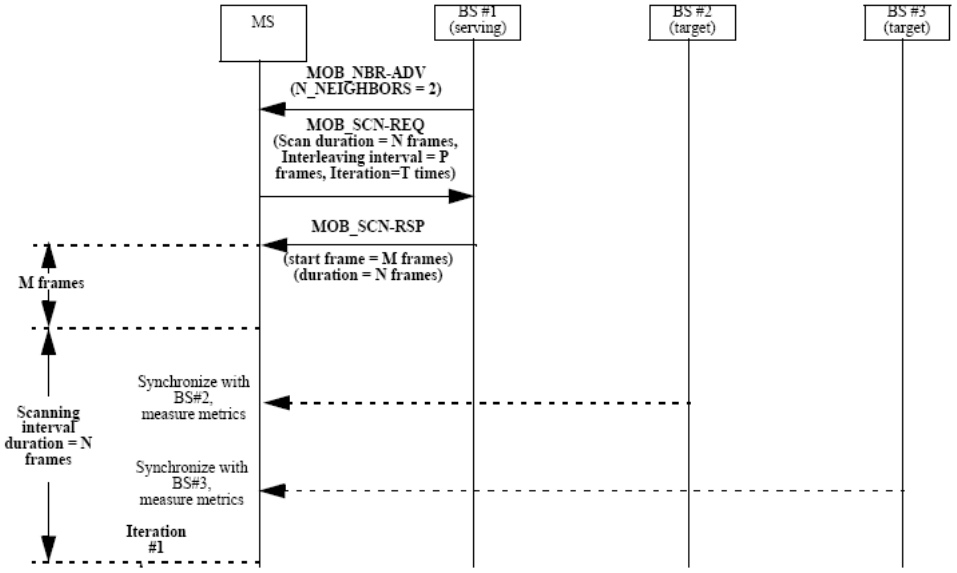

Figure 2.4: An example neighbor advertisement and scanning procedure [IE[3]05].

The actual cell re-selection is started when the MS begins to scan potential TBSs in order to find out if they can offer sufficient signal strength and quality (see Figure 2.4). The MS has to request permission from the Serving BS for the scan with a MOB_SCN-REQ message. The SBS responds with a MOB_SCN-RSP message defining a period of time when it will buffer the traffic sent to the MS, enabling

[4]A third reason to do a handover is to force an MS that is causing a lot of interference in one frequency band to handover to another frequency band (a so called confinement handover [Mou92]).

the MS to scan the TBS. Such scanning can therefore endanger the QoS of delay sensitive service flows.

The MS can also create a preliminary association to the TBS by conducting initial ranging. Such association enables the MS to acquire and record more detailed ranging parameters already in the cell reselection phase to speed up the potential upcoming handover to a TBS. The access network can participate in the association process with different levels, depending on how high guarantees of the success of ranging and the length of the ranging process is wanted. The chosen association level will be sent in the MOB_SCN-REQ message to the SBS and relayed to the TBS through the access network.

Scanning can be conducted also during the handover decision and initiation phase if handover initiation takes a long time and the original scanning results might be outdated.

2.1.3.2 Phase 2 for an MS initiated rescue handover

If the handover is an MS initiated rescue handover, the MS will first trigger the scanning procedure described above, when the signal quality has deteriorated to a value under a predefined threshold. Based on the measurements made of the TBS signal strength and quality, the MS will decide whether it should actually handover to the TBS. There are many algorithms that could be used for the decision and one commonly known example is the relative hysteresis algorithm [Gud91][Pol96] which eliminates the "ping-pong" handover effect, that results from premature reaction to quickly changing channel conditions[5].

If the condition for a handover is fulfilled the MS will start the handover by sending a MOB_MSHO-REQ message[6]. In this message the MS will include a list of TBSs it would like to handover to.

If an MS initiated network controlled handover is conducted, the Serving Base Station will first send a message through the backbone network to the Target Base Stations inquiring if they have sufficient amount of resources to admit the MS. The TBSs will respond by indicating the kind of QoS they can provide. This is the most interesting part of the rescue handover for us, as we are mostly concerned with resource utilization.

The list of TBSs that have sufficient resources will be sent to the MS in a MOB_BSHO-RSP message. The MS will make a final indication to which TBS it has chosen to

[5]The MS can also decide to do a handover in the case of overload in the BS. This would be done especially with MSs that have more intellect and only BE connections.

[6]In other systems such as GSM a rescue handover is initiated by the access network and only assisted by the Mobile Station (Mobile Assisted Handover (MAHO)). BS controlled rescue handover is also possible in Mobile WiMAX.

perform a handover to with a MOB_MSHO-IND message. The MS can cancel the handover, in case the TBS signal quality has dropped, by stating the BS id of the Serving Base Station in the MOB_MSHO-IND message.

In an uncontrolled handover the MS will just try to enter and range the TBS without any prior signaling with the SBS or backbone negotiation.

2.1.3.3 Phases 1 and 2 for a BS initiated directed handover

The traffic load unbalance in the system will trigger BS initiated directed handovers. The logic behind the handover triggering is mostly an issue of the access network and it will be discussed in detail in the following chapters.

What really becomes an issue, in terms of the radio link, is when and how scanning should be conducted in relations to the handover decision. Scanning is basically the only way we can find out which MSs are in the overlapping areas of two Base Stations. Roughly said there are three options: to do the scanning before the decision, after the decision or in a hybrid way both before and after the decision.

If scanning would be done before, a list of MSs residing in the overlapping areas would be kept by sending unsolicited MOB_SCN-RSP messages telling the MSs to scan the TBSs defined in the MOB_NBR-ADV message. This would result in a high number of periodically occurring scans. This could be mitigated by narrowing down the number of candidate MSs in the overlapping areas. Location estimation algorithms could be used and the list could be kept only of static MSs.

Doing the scanning after the directed handover decision would have the advantage, that periodical scanning to maintain the list of MSs in overlapping areas would not have to be done. The drawback is that it might take too long to resolve which MSs reside in the overlapping areas after which it might already be too late to conduct load balancing. Scanning after the handover decision would however be required, in the case where the MS wants to associate to the TBS before disconnecting from the BS to enable a more reliable handover execution.

Irrelevant of how scanning and association is conducted, the directed handover will be initiated with a MOB_BSHO-REQ message where the Serving Base Station will recommend a TBS that the MS should handover to. As in the MS initiated handover, here the MS will also make a final indication that it is about to perform a handover with a MOB_MSHO-IND message. The MS can also do additional scans (by its own initiation or BSs) to make sure that the signal strength and quality of the TBS are still sufficient.

Because scanning after the handover decision seems necessary, the best choice would be either to do scanning only after the decision or both before and after.

2.1.3.4 Phases 3 and 4

As stated before the remaining phases 3 and 4 are more or less the same irrelevant of whether the handover is an MS initiated rescue or a BS initiated directed handover. In terms of QoS the goal in phases 3 and 4 of the handover is to do them as quickly and reliably as possible because in the case of a normal Hard Handover (HHO) execution the MS connection will be interrupted.

Association to the TBS, MS context transfer between the SBS and TBS and different fast handover mechanisms (e.g. Fast Base Station Switching (FBSS) and Macro Diversity Hand Over (MDHO)) are ways to enable fast handover execution (and interruption time) and can be used to guarantee the delay QoS demands of real-time service flows. However the shorter interruption time and reliable handover execution comes with the prize of larger handover signaling overhead and waste of resources.

Optimization techniques for Hard Handovers

In the synchronization to the Target BS downlink phase (phase 3), an MS will synchronize to the DL transmission of a TBS and obtain DL and UL transmission parameters. As mentioned before, if an MS had previously received a MOB_NBR-ADV message including the DCD and UCD messages, this process may be shortened.

In fourth phase of the handover, the ranging and network re-entry will be done. The duration of ranging depends on the association level the MS has used to associate to the TBS in the cell reselection phase. If no association was done (Scanning without Association) ranging might take a long time since ranging parameters have to be acquired. If association has been conducted the TBS can use a Fast Ranging Information Element (Fast_Ranging_IE) in the UL-MAP to provide an MS with a contention or non-contention based initial ranging opportunity.

If association with level 0 (Scanning or association without coordination) is used, ranging is shortened because the initial ranging parameters will already be available, but with this level the TBS won't have any prior knowledge of when the MS will be ranging and will therefore offer contention based ranging.

Both in association level 1 (association with coordination) and 2 (network assisted association) a non-contention based initial ranging opportunity is offered. The difference between level 1 and 2 is that with level 1 a regular non-contention based ranging opportunity is used whereas with association level 2, a dedicated Code Division Multiple Access (CDMA) ranging code will be reserved for the MS and network assistance will be offered already in the cell reselection phase making it more reliable.

In fourth phase of the handover, re-registration, authentication and negotiation for basic capabilities will also be done. If the SBS and TBS are able to exchange the

context information of the MS through the backbone (e.g. service flow parameters), some or all of the re-entry procedures can be skipped. This kind of hard handover optimization (Optimized Hard Handover (OHHO)) will shorten the duration of network re-entry dramatically[7].

Fast handover schemes

Where in Hard HO (normal (HHO) and Optimized Hard HO (OHHO) the MS disconnects from the serving BS before connecting to the Target BS and has no connection during phases 3 and 4, with fast handovers schemes Fast BS Switching (FBSS) and Macro Diversity Handover (MDHO), the MS is connected to one or more BSs during the handover execution. In other words during handover execution in FBSS and MDHO, phases 3 and 4 are repeated for many BSs making it a quite heavy procedure.

In FBSS (also known as a seamless handover) the MS maintains a list of active BSs that it has established a connection to. The MS is able to receive and transmit every frame from any BS within this "Active set" and therefore no handover interruption should occur [Bec06].

MDHO takes this one step further. With a MDHO handover the MS is able to communicate with all of the BSs[8] enabling diversity combining to be used to get the optimal signal quality in both downlink and uplink.

The question of how many Target Base Stations are included in the FBSS and MDHO handovers is an interesting one. In [Cho05], it was proposed to associate only one Target BS instead of several Target BSs to reduce the time of network acquisition and active set updates. Also when examining probable cell layout scenarios, it seems more likely that the overlapping areas will be formed as a result of the coverage of two to three Base Stations (see Figure 2.2) and hence the size of an active set would be limited to two or three.

Using the handover execution mechanisms

We have seen that there are many different ways the handover can be executed. Such variety can be used to meet the needs of different types of traffic conducting a handover. In [Don07] it was proposed to map the association levels, FBSS and MDHO mechanisms to the different Scheduling services. In [Bec06] the possible usage of these schemes in terms of mobility was described. Both of these are summarized in Table 2.2.

[7]WiMAX Forum has developed these optimization techniques with a goal of keeping Layer 2 handover delays to less than 50 milliseconds [WiMAX].

[8]In MDHO the list of Base Stations is called a Diversity Set

Table 2.2: Example usages of different handover execution mechanisms [Don07] [Bec06].

Mapped to Scheduling Services	
HO execution mechanism	*Scheduling Service*
Ass. Lev. 0, 1	BE
Ass. Lev. 1, 2	nrtPS
Ass. Lev. 2, FBSS	ertPS, rtPS
FBSS, MDHO	UGS
Mapped to Mobility	
Mobility	*HO execution mechanism*
Portability Low walking speed	HHO, OHHO (Ass. 0, 1, 2)
Simple mobility Low vehicular speed	HHO, OHHO (Ass. 0, 1, 2)
Full mobility High vehicular speed	FBSS, MDHO

The question that is of importance to us, in terms of load balancing, is how does the usage of these handover execution mechanisms affect resource usage during handovers. From the list we can see that FBSS and MDHO will mostly be used for MSs with high mobility and they are therefore more likely to occur in rescue handovers. However we can also see that directed handovers with FBSS and MDHO are possible especially with real-time services.

With hard handovers the situation is straightforward because resources are used at one BS at a time. With FBSS resources are used in the same way as with hard handovers but in addition each BS in the active list has to reserve resources to be able to accept transmission from the MS in any frame during the handover execution. This could be therefore an issue also in load balancing.

MDHO is trickier because it will also use the resources of the BSs in the diversity set. It would have some sort of a temporary effect in the load situation. Predicting the side effects that load balancing initiated directed handovers using MDHO would cause to the whole system is very difficult. The effect it has on load balancing could be target of future research.

All in all the additional resource usage of FBSS and MDHO becomes an issue only when the handover execution lasts for a long time and the load level in all TBSs is high and free resources scarce. Still it might be beneficial to minimize the number of load balancing triggered directed handovers especially with non-real time services, due to the heavy signaling and execution procedures that might also degrade QoS.

As can be seen from this subsection, there are many ways to execute a handover after the decision to make the handover has been made. This is useful background information for us, but as the purpose of this thesis is not to investigate different ways handover execution and interruption time could be optimized, but to examine Mobile WiMAX handovers on the point of view of efficient utilization of system wide Radio Resources, the issue won't be covered anymore.

2.2 Overview of the WiMAX Forum Access Network Architecture

The job of the WiMAX Forum is to design and enforce a Network architecture that will guarantee the interoperability of the WiMAX products of different vendors. This end-to-end network architecture is currently described in two documents: WiMAX Forum Network Architecture Stage 2 [ASN2] that describes the general framework of the network and Stage 3 [ASN3] that specifies detailed procedures and messages of the network. The documents describe the Network Architecture in terms of mobility, security, authentication and inter-networking with other systems such as 3rd Generation Partnership Project (3GPP) based systems [Lax06].

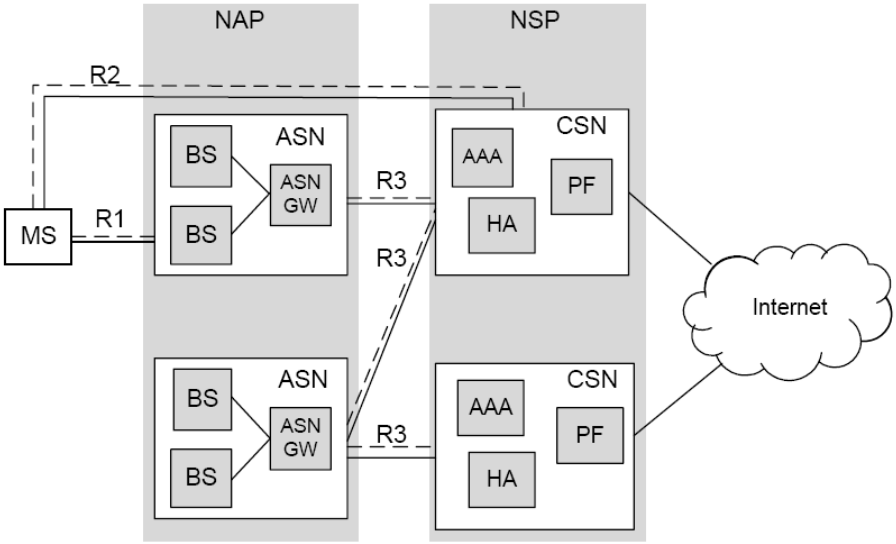

Figure 2.5: WiMAX Forum Network Architecture [ASN2].

The Network Architecture consists of two business entities: a Network Access Provider (NAP) which governs a set of Access Service Networks (ASN) and a Network Service Provider (NSP) which governs Connectivity Service Networks (CSN) that include Authentication, Authorization and Accounting (AAA), Policy Function (PF), Home Agent (HA) and other required functionalities.

Since we are mostly interested in the system wide Radio Resource Management of a cluster of Base Stations we will limit ourselves to examining only the access network part (ASN) of the network architecture. In the following subsections we will take an overview look at the ASNs RRM framework in terms of its topology and handover support. To further limit our scope, from now on we will only concentrate on the intra-ASN case of handovers excluding both inter-ASN and inter-technology handovers.

2.2.1 ASN network topology

The Access Service Network (ASN) is a key network element in the WiMAX Forum Network Architecture and it consists of one or more Base Stations and ASN Gateways. The tasks of the ASN include connection establishment between the MS and the ASN, Radio Resource Management, paging and location and mobility management between Base Stations. The ASN manages these only on radio link (MAC) level and leaves most of the higher level management to the other network entities. This makes the deployment of individual access networks possible.

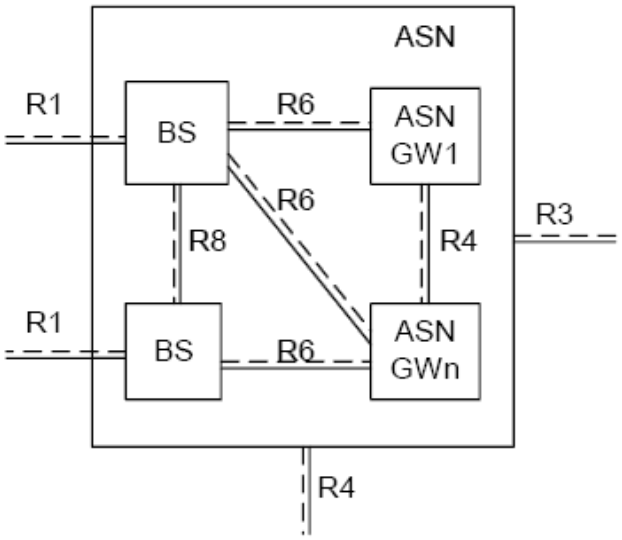

Figure 2.6: The ASN reference architecture [ASN2].

The most important aspect that we want to investigate is what kind of system wide RRM support does the ASN provide in terms of load balancing and traffic prioritization and the kind of handover framework it gives.

2.2.1.1 ASN RRM functional entities

The ASN includes an RRM architecture that enables efficient radio resource utilization in the WiMAX network. The RRM procedures in the ASN can be used for decision support in admission control for new flows and rescue handovers, triggering load balancing and handover preparation and control.

RRM is composed of two functional entities Radio Resource Agent (RRA) and Radio Resource Controller (RRC) that handle RRM messaging within the ASN (for their location in different ASN profiles see Figure 2.7).

Radio Resource Agent (RRA)

The RRA resides in the BS and has three main tasks: it maintains a database of collected radio resource indicators of the MSs registered to it, communicates with the MSs and the RRC and is responsible for assisting local Radio Resource Management in decision making.

The database of collected MS radio resources indicators may include current physical service level (channel bandwidth), error rates and available radio resources which are utilized to form the RRM content of the control signaling messages.

The RRA reports its own radio resource status with these messages to the RRC and receives updates from the RRAs located in other BSs from the RRC. The RRA also uses some of this received information to create messages, such as the neighbor advertisement messages discussed earlier, sent to the MS through the air interface.

In addition, RRA controls the local radio resources of the BS it belongs to based on measurement reports from its host BS and also based on radio resource usage information of the other BSs received from the RRC. The tasks of the RRA include among others local power control, service flow admission control and load balancing control which will initiate the directed handovers. What is especially interesting to us is load balancing control and service flow admission control procedures relating to rescue handover and traffic prioritization.

Radio Resource Controller (RRC)

The RRC can be located in the Base Station or in the ASN-Gateway (ASN-GW) node depending on the ASN profile. The main responsibility of an RRC is to collect radio resource indicators from associated RRA(s). In other words it is in charge of communication between and across RRAs and can terminate and combine messages containing the information of individual BSs to an aggregated status update.

In the case where the RRC resides in the BS, there is the possibility to have an RRC relay in the ASN-GW for the purpose of relaying the RRM messages. It will

however only have the relaying functionality of an RRC and can not terminate the RRM messages[9].

2.2.1.2 ASN profiles

Different ASN Profiles have been specified in Mobile WiMAX to offer a wide range of access network deployment possibilities. Such flexibility brings additional value to the WiMAX system but raises questions on what should actually be implemented. Especially the BS vendors seem to be concerned that they will be forced to implement every protocol option in the specification [Li06]. There are naturally many questions also within the WiMAX Forum on the technical and business merits of each network profile [Hu07].

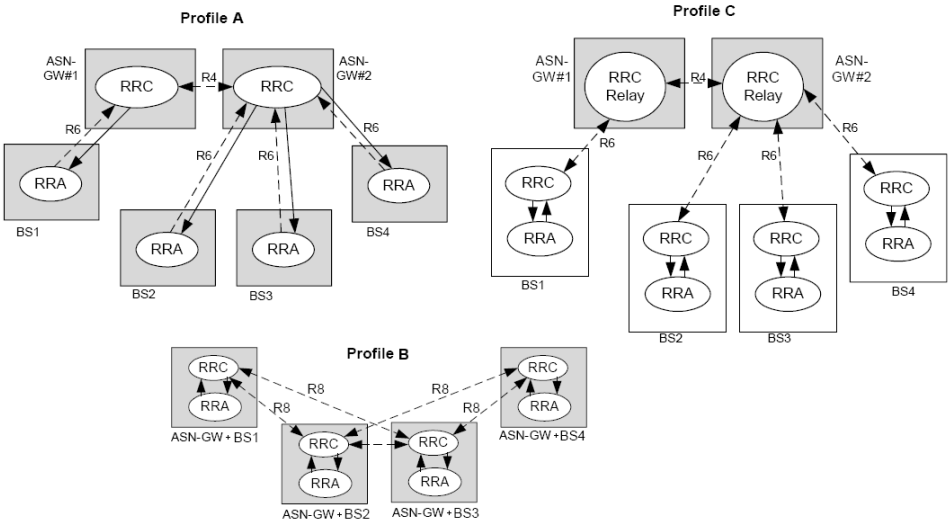

Figure 2.7: ASN profiles [ASN2].

Three ASN profiles are offered and they are determined by the location of the RRM functional entities RRC and RRA discussed above. Profile A and C feature a hierarchical structure with separated ASN-GW and BS nodes, where in profile A the RRC is located in the ASN-GW providing a more centralized model and in profile C the RRC is co-located with the RRA in the BS providing a more distributed solution for RRM. Handover control is divided in the same manner for profile A and C, meaning that in profile C only the non-mobility related tasks are performed in the ASN-GW [Hu07]. Profile B offers a totally distributed flat model with no hierarchy, where both BS and ASN-GW are implemented as a single node

So what are the different advantages and disadvantages of these profiles? Profile A has the advantage that it reduces backhaul signaling, because centralized

[9] All in all the name Radio Resource Controller might seem a bit misleading because local radio resource controlling is actually conducted by the RRA.

RRC aggregates control messages. Profile A enables also the possibility to conduct macro level diversity combining making a soft handover (MDHO) possible. The main disadvantage with profile A is that it makes interoperability between a BS and a ASN-GW from different vendors difficult and therefore limits scalability. As a result fewer vendors are interested in profile A.

All in all profile A relates more to the access network architectures used in today's cellular networks (GSM, UMTS) where most of the intelligence and control workload are in the gateway nodes (Base Station Controller (BSC), Radio Network Controller (RNC)) closer to the core network. Profile B provides a simple flat architecture and is therefore well suited for small scale and can be very expensive in large scale deployment.

The main advantage of profile C, especially in relations to profile A, is that it offers good interoperability between the BS and ASN-GW and enables the possibility to get both from different vendors. This results in good scalability. The drawback is extra backhaul signaling.

All in all it seems that in Mobile WiMAX most of the RRM intelligence and handover control will at least in the early stages reside in the BS[10]. The different profiles will mostly just change the way the BSs communicate with each other, but the same information will still be available for all. So irrelevant of the way this messaging is done, load balancing and handover prioritization would be initiated and controlled in the BS.

Such an approach makes sense if we compare the traditional cellular systems to the philosophy that the Internet introduced. The Internet brought forth the concept of having the network control intelligence in the terminal instead of the approach used for example in cellular networks, where the intelligence is closer to the core network (BSC, RNC). Having the intelligence in the terminal eases network operation but makes it more difficult to guarantee QoS. As Mobile WiMAX is more terminal driven the access network solution it provides, seems to be a compromise between these two approaches resulting in a unique way to offer both a possibility for differentiated QoS and easy operational maintainability.

As profile C seems most likely to be deployed by many vendors we will from now on concentrate on that, but will still try to design the schemes to be such that they could be deployed with all profiles.

[10]Centralized RRM could be an extension to profile A.

2.2.2 Handover and RRM procedures

So what kind of RRM messages will be sent in the ASN in the case of handovers? As stated before handovers are on the other hand a tool for us in the form of load balancing triggered BS initiated directed handovers and on the other hand a problem in the form of MS initiated rescue handovers. What is especially interesting to us is how ASN signaling is conducted in these cases and what are the contents of these messages.

2.2.2.1 Handover procedures

The WiMAX Forum network architecture divides the handover process to two phases: handover preparation phase and handover action phase. When comparing these two to the handover phases defined by the IEEE 802.16e standard, described in subsection 2.1.3, we see that the handover preparation phase corresponds to handover decision and initiation and that handover action phase corresponds to synchronization to the TBS downlink and network re-entry (including ranging).

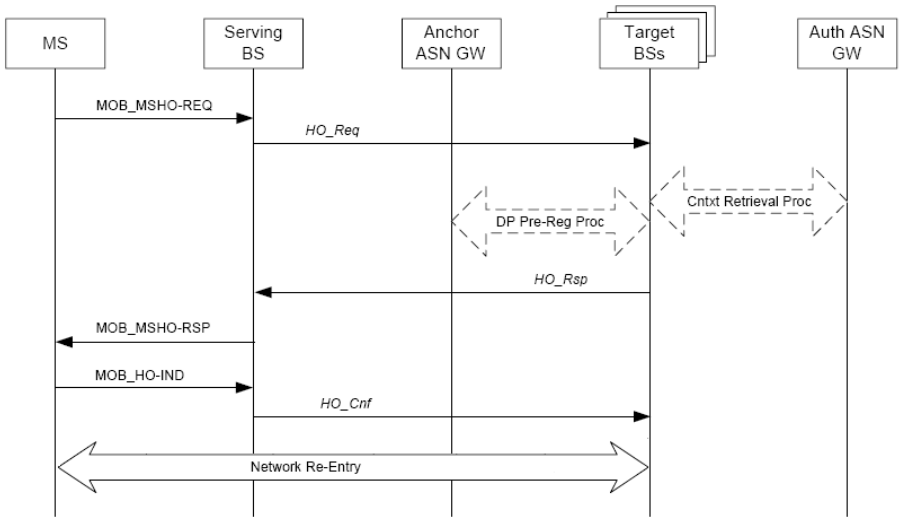

Figure 2.8: MS initiated handover [ASN2].

When an MS initiated handover is triggered the MS sends a MOB_MSHO-REQ message to the Serving BS defining the set of TBSs it would like to handover to. This begins the handover preparation phase between the SBS and the TBSs. The SBS will first send a HO_req message to inquire from the candidate TBSs if they have enough capacity to receive the MS. If yes, the TBS will reserve resources for the MS and can possibly even conduct preliminary context retrieval and datapath registration to enable fast execution during the handover action phase. The TBS will respond with a HO_rsp message indicating if it has enough resources to accept the MS and informing also the QoS level it can provide. The SBS will send the remaining pruned candidate TBS set to the MS in a MOB_MSHO-RSP message,

based on which the MS will decide where it will handover to.

The MS will send a MOB_HO-IND message to indicate its final decision which initiates the action phase. The selected TBS will be informed of the incoming MS with a HO_conf message with a "confirm" value, so that the TBS has time to complete datapath registration and other required tasks. Also the context information of the MS can be exchanged between the SBS and the selected TBS to enable even faster handover execution during network re-entry procedures. HO_conf message with value "rejected" will be sent to the other candidate TBSs so that they can release the reserved resources.

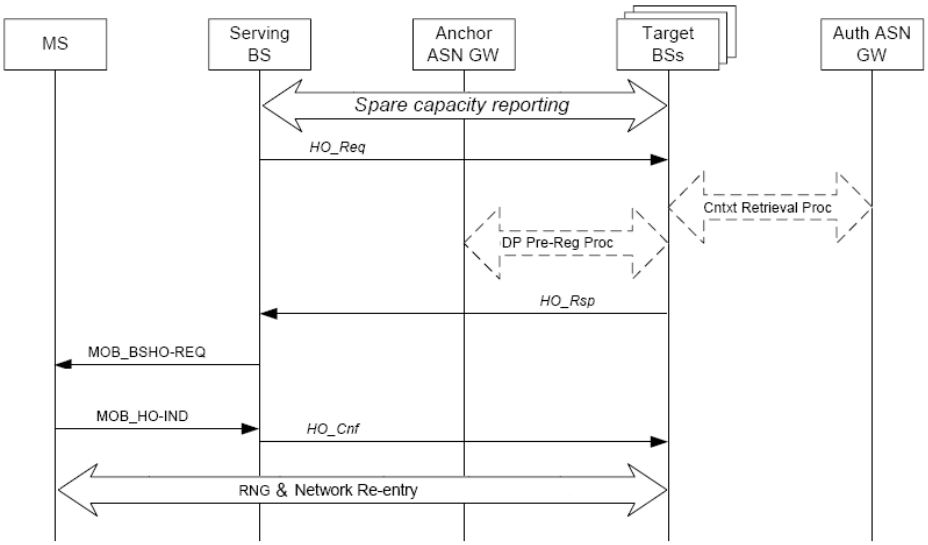

Figure 2.9: BS initiated handover [ASN2].

The BS initiated directed handover preparation phase is preceded by a procedure where load information of all the BSs is exchanged in the form of spare capacity reports. There reports are used to determine when the BS is overloaded and trigger load balancing with directed handovers. If a BS is overloaded and the MSs that reside in overlapping areas are known, the directed handover procedure for a single MS can be initiated by the BS by sending HO_req messages to the other less congested TBSs that cover the overlapping area. From there on the preparation phase and the action phase of the handover continue as in the MS initiated case.

2.2.2.2 Framework for load balancing

What kind of a framework does Mobile WiMAX offer for load balancing? The specification defines many RRM messages that are used in the communication between the RRCs and the RRAs. The core load balancing tool in the ASN is the Spare

Capacity Reporting procedure. This procedure can be used to keep all the Base Stations in the ASN up to date of the resource usage of their peers. A procedure to report physical RRM parameters is also provided and can be used as supplementary information for making the load balancing decisions.

Spare Capacity Report (SCR)

The spare capacity reporting procedure, preceding the BS initiated directed handover, has many similar qualities as the scanning procedure conducted in the MS initiated rescue handover. In both cases a handover is conducted as a result of a lack in resources, but where signal quality and strength is compared in the rescue handover, resource utilization is compared in a directed handover.

The Per-BS Spare Capacity Report (SCR) can be sent by request or configured to be sent periodically after passing a threshold in resource usage. Spare capacity is described with the Available Radio Resources-indicator. It describes the average number of available slots per frame. The value is averaged over a predefined interval[11] and given in percentages.

The spare capacity is reported for both UL and DL and is defined as the *set of free slots not used by any non Best Effort service flow class*. The reason for omitting the resource usage of BE service flows apparently lies behind the idea that MSs with only BE services should conduct and trigger load balancing themselves. The SCR also includes a Radio Resource Fluctuation field that describes traffic fluctuation. The field describes the degree of fluctuation in channel data traffic throughputs for the Base Station and is dependent on the variability of the served traffic and channel fluctuations. This value could be used to eliminate unnecessary load balancing handovers. Based on this information the BS can make the decision whether to trigger load balancing.

Physical Parameters Report

The Physical Parameters Report is conducted per MS and is done by request. The report includes the Carrier to Interference and Noise Ratio (CINR) and Received Signal Strength Indicator (RSSI) for both UL and DL for the given MS. It also includes the Physical Service Levels (PSL) for both UL and DL which describe the channel rate available for the MS. The PSL value corresponds to the MCS that can be used with the MS under the channel conditions. This information can be utilized when choosing the optimal TBS from the set of candidate TBSs and could also possibly be used in the process where the SBS tries to find out which MSs are in the overlapping areas.

[11]The interval is 200 frames by default, which corresponds to one second if frame length is 5 ms.

Spare Capacity Report per QoS profile

The specification defines also a way to report Spare Capacity per QoS profile. The basic idea is that a TBS could pre-calculate the number of MSs, with a specific QoS profile and PSL level, it could admit. The calculations would be based on the current channel and traffic conditions. This information could enhance the accuracy of the load balancing decisions.

However in release 1 of the WiMAX Forum Network Architecture, Spare Capacity report per QoS profile will not be used directly for load balancing[12]. In release 1 it can only be used in the handover preparation phase in the HO_req and HO_rsp messages. In other words, in the case of load balancing with directed handovers, it could be utilized to decide how many handovers should be conducted to each lightly loaded TBS.

2.2.2.3 Resource reservation for handover connections

Be it an MS initiated rescue or a BS initiated directed handover, after a decision to do a handover has been made, resources have to be reserved from the candidate Target Base Stations.

The resource reservation procedure in the TBS is controlled with the HO_req and HO_rsp messages mentioned above. When sending the HO_req message to the TBS, the SBS specifies the QoS parameters guaranteed for the service flows of the MS. After receiving the HO_req, the admission control function of the TBS will calculate whether it has enough resources to admit the connection and if not it will indicate failure due to insufficient resources. The TBS may also suggest a lower QoS profile in the HO_rsp message.

If no prioritization of handover calls is done, the same admission control procedure used for new service flows, will be used for the handover service flows. As stated earlier in the case, where none of the candidate TBSs have sufficient resources, this will result in a handover drop. Prioritization of handover calls could be implemented by keeping a guard band for the rescue handovers coming into the TBS. It would mean that if resource reservation[13], would go over this guard band limit, admission control would start to block new calls but would still accept rescue handovers.

[12]It will be supported in later releases.
[13]The minimum resources reserved for the service flows.

Chapter 3

Background Research

Having gone through the key system aspects of IEEE 802.16e and WiMAX Forum network architecture, we can start considering how the actual load balancing with handovers could be conducted. Also we can start to examine the handover and traffic prioritization aspect in a more concrete way to get an understanding of its role.

A good place for us to start is to see what kind of prior research has been conducted regarding these issues and to examine how the existing ideas could be applied to Mobile WiMAX.

3.1 Load balancing with handovers

Load balancing with handovers will be the way system wide load balancing will be conducted in Mobile WiMAX. In this section we will first classify the different ways load balancing can be conducted in telecommunications systems, see what kind of a relationship Mobile WiMAX has with them and finally dig into the theory behind load balancing from the point of view of Mobile WiMAX.

Secondly, having gained a good understanding of the balancing method used, we will do a literary review of the previous research conducted and do some initial speculations of their feasibility to be applied in Mobile WiMAX.

3.1.1 Introduction

Here we will first briefly discuss how load balancing can be conducted with resource allocation and load distribution based schemes and see how and if they could be used in Mobile WiMAX. Then we will move on to study the theory behind load distribution based load balancing (that will be used in Mobile WiMAX) and discuss what kind of a load metric should be used and how load measurement is conducted.

3.1.1.1 Classification

Load balancing schemes that try to solve the hotspot problem can roughly be divided to resource allocation schemes and load distribution schemes [Kim07].

Resource allocation schemes

The idea behind balancing the system load with resource allocation is to *bring the resources (unoccupied frequencies) to where most of the users are located.* In resource allocation schemes, a centralized element allocates additional resources to hotspot cells. One example of this is channel borrowing where a congested Base Station can borrow the channel of lightly loaded Base Stations.

Channel borrowing requires that the system supports Dynamic Channel Allocation (DCA), which is an enhancement to the traditional Fixed Channel Allocation (FCA). DCA is able to adjust to changing traffic whereas FCA will keep the same frequency assignments irrelevant of the traffic load [Ira00]. Although Mobile WiMAX provides a flexible way to allocate frequency resources making DCA between BSs possible, DCA won't be used at least in the early stages of deployment. FCA will be applied for the frequency sets resulting from PUSC sectorization.

Load distribution schemes

Where in resource allocation based load balancing the aim is to bring the resources to where most of the traffic is, with load distribution the goal is to *direct the traffic to where the resources are.* The way to do this is to use handovers.

Load distribution with handovers can be conducted many ways. One commonly used simple approach is *cell breathing.* There load balancing is done by adjusting the transmission levels of the SBSs pilot signal (shrinking the cell) according to the traffic level, resulting in a situation where MSs at the edge of the cell are forced to conduct rescue handovers. In systems based on CDMA cell shrinking happens automatically as the number of MSs increases.

This approach could be used for load balancing also in Mobile WiMAX. However there are some disadvantages [Lee07]. The biggest drawback would be that a BS would have less control on where and when an MS would conduct the handover and hence the possibility to guarantee QoS system wide would decrease. At worst the MS forced to initiate a rescue handover might not have any other TBSs in range and the connection would be dropped.

Another method for load distribution is *traffic load based MS initiated handovers.* In this approach the load balancing logic resides in the MSs. It is already in use for example in some WLAN terminals which can choose the least congested Access Point (AP) based on measurements made of the candidate APs. MS initiated

load balancing handovers can be easily conducted in Mobile WiMAX based on the available resource information broadcasted in the MOB-NBR_ADV message. It will be used in Mobile WiMAX at least for MSs that have only BE service flows but possibly with other Scheduling Services as well[1].

The load distribution method most important for us, is the *directed handover* where the congested SBS forces the MS to handover to a less congested TBS. This is a very good approach for Mobile WiMAX because it enables better control for the BS and therefore makes it possible to guarantee QoS system wide.

From now on we will concentrate only on load balancing based on BS initiated directed handovers. All in all the load balancing effect of load distribution is highly dependent on the size of the overlapping area between the Base Stations and therefore in some situations it might not be able to release enough resources to fulfill all QoS guarantees. This is why the handover guard bands and other traffic prioritization schemes that will be examined in section 3.2 are quite likely needed also in Mobile WiMAX.

The usage of relay stations will be introduced to Mobile WiMAX in the future with IEEE 802.16j. It will provide broader overlapping areas and will even enable the possibility to dynamically direct the coverage of the relay stations to the congested areas. It could therefore be characterized as another resource allocation scheme because it would bring resources to the hot spot cell. In [Yan05] the effect of relay stations was examined and it was shown that with relay stations load balancing could be so effective that even handover prioritization would not be a critical issue anymore.

3.1.1.2 Theory

Load balancing can be defined as the process of dividing and distributing workload (jobs) between many processors (servers) so that more workload can be served. Load balancing has been mostly used in computer systems for load sharing, but has also been applied in telecommunication.

In the case of a single Base Station, the packets of the service flows would correspond to the jobs to be processed, and the Base Station would correspond to a processor that serves them. Each BS (more specifically the scheduler in the BS) could be modeled on the packet level with two (one for UL and DL) non-preemptive G/G/1-priority queues if all users are aggregated to a send. On the other hand each MAC slot could also be viewed as a server.

[1] One interesting idea to enhance the load balancing support by the MS was presented in [Kim07]. The basic idea is to delay rescue handovers to hotspot BSs and to speed up rescue handovers to lightly loaded BSs.

In general, load balancing can be conducted in a static or dynamic manner. Static load balancing is independent of the state of the system where as in dynamic load balancing, decisions are made based on the current loading situation and availability of resources. Load balancing can also be done in a distributed or centralized way. The centralized approach reduces signaling but is sensitive to node failure. The distributed approach on the other hand is simple and robust but requires a great amount of signaling and cannot optimize the system in the same way as the centralized approach does. When applied to WiMAX, as discussed earlier, the most likely choice is to use dynamic load balancing in a distributed manner but the centralized approach is also possible with ASN profile A [Wu05].

Important elements in terms of load balancing are *load metric*, *load measurement* and *load balancing operation* [Wu05]. We will discuss load balancing operation, which basically specifies how the load balancing is triggered and executed, in detail later. In the following we will take a look at how the load balancing metric could be defined and how the load could be measured in Mobile WiMAX.

Load balancing metric

The load metric should describe well the loading situation in relations to the usage of common resources. The shared resources in the radio link of an OFDMA system can be divided to time, frequency and power. The usage of these resources depends on the transmission power and the MCS[2].

Commonly used load metrics are number of calls and blocking probability in traditional cellular networks and packet loss, throughput and delay in wireless networks such as WLAN. Measuring users or connections in Mobile WiMAX is inaccurate because MSs might have many service flows and furthermore each service flow might have different characteristics. Throughput does not consider what MCS is used and therefore it cannot be known when the maximum resource utilization has been reached. Packet loss and delay only give indirect information of the loading situation and should therefore be only used for decision support.

The basic resource measurement unit in Mobile WiMAX is one slot. This is a good and accurate indicator of resource utilization because it describes the resources not just in terms of throughput, but also in relations to the MCS used and therefore also takes into consideration the channel conditions. It has been a natural choice for load metric also in the WiMAX Forum network architecture.

[2]In this report we will assume that network dimensioning and power control is conducted so that power won't be a critical issue. The effect of power in resource utilization could however be the target of further research.

Load measurement

The most simple way to evaluate how balanced the system is, is to calculate the average load of the whole system [Vel04]

$$L = \frac{\sum_{i=1}^{n} U_i}{n} \tag{3.1}$$

where n is the number of Base Stations, and U_i is the resource utilization of Base Station i and compare this average to the individual resource utilizations U_i of each Base Station. To describe the loading state of the whole system with one value the following balance index has been defined [Jai84][Chi89]:

$$\beta = \frac{(\sum_{i=1}^{n} U_i)^2}{n(\sum_{i=1}^{n} U_i^2)} \tag{3.2}$$

The index β gives a value between 0 and 1, where 1 indicates that the system is balanced. Since it is quite likely that the uplink and downlink subframe division in Mobile WiMAX will be static we will define resource utilization of a Base Station as

$$U_i = \max\Big(U_{DL,i}(\mathbf{A}), U_{UL,i}(\mathbf{A})\Big) \tag{3.3}$$

where $U_{DL,i}(\mathbf{A})$ and $U_{UL,i}(\mathbf{A})$ are the resource utilizations of downlink and uplink subframes with a given association matrix $\mathbf{A}$. The association matrix $\mathbf{A}$ describes to which BS each MS is associated to, with $a_{i,j} = 1$ indicating that MS j is associated and $a_{i,j} = 0$ indicating that MS j is not associated to BS i. The set of possible $a_{i,j} = 1$ values is limited to the BSs covering the overlapping area where the MS resides.

The resource utilization of the downlink subframe for BS i can be calculated as

$$U_{DL,i}(\mathbf{A}) = \frac{\sum_{j=1}^{k} (\frac{B_j^{DL}}{c_{i,j}^{DL}} a_{i,j})}{U_{DL,tot} S_{fps}} \tag{3.4}$$

where k is the total number of MSs in the system, B_j^{DL} is the total throughput of all the service flows in the downlink for MS j. $c_{i,j}^{DL}$ is the number of bits carried per slot in the downlink based on the MCS used between MS j and BS i in the downlink, $U_{DL,tot}$ is the total number of slots in the downlink subframe and S_{fps} is the main frame rate. The uplink subframe resource ulitization $U_{UL,i}$ is defined in a similar manner. As can be seen the final resource utilization is given as percentages of the total number of slots as defined in WiMAX Forum Network Architecture.

The evident problem that arises when measuring the load of the system is the fluctuation of both traffic B_j^{DL} and the channel $c_{i,j}^{DL}$. Doing load balancing handovers prematurely as a reaction to these variations might cause a similar "ping-pong" phenomenon as in the rescue handover decision, if a relative hysteresis margin is not

used. Conducting many unnecessary handovers would be especially bad for real-time service flows for which the handover process can be very heavy as discussed earlier. This introduces a tradeoff between how much unbalance is tolerated and how many load balancing triggered directed handovers are conducted [Vel04].

3.1.2 Previous research

We concluded in the previous subsection that BS controlled load distribution will most probably be the way load balancing is done at least in the early stages of Mobile WiMAX. The aim is therefore, in the case of congestion, to direct connections to where there are free resources. There are two main schemes relating to this, that have been the target of research: *Directed Handover* and *Directed Retry* [Ira00]. In this subsection, we will take a look at the research conducted for both and do some initial considerations on their feasibility for Mobile WiMAX.

3.1.2.1 Directed Handover

Load distribution with Directed Handovers (DH) is one of the most potential methods to conduct load balancing in Mobile WiMAX. It has been previously studied with traditional cellular technologies such as GSM, with WLAN Acces Point clusters and also to some degree in terms of IEEE 802.16e networks. In terms of Load balancing operation the most important questions that we would like to answer are *when* load balancing with directed handovers should be initiated and *how* and to *where* the directed handovers are actually conducted.

Directed handovers in a traditional cellular network

In [Fuj92] cell load triggered directed handover mechanisms were considered. In one of the proposed schemes, in the case of overload in an SBS, the MSs with the longest radio distance (worst signal strength and quality) are handed over to less congested TBSs. The simulations revealed that by using the mechanisms, traffic performance can be improved by about 50 % under the condition that offered traffic is uniformly distributed. The paper also considers the resulting interference effect of a directed handover and proposes a method where only the handovers that don't result in excessive co-channel interference for the network are actually performed. Only one fixed threshold is used to trigger load balancing. This might work well with traditional cellular networks with rather fixed traffic and Modulation and Coding Schemes, but using it with Mobile WiMAX where both traffic and MCS vary to a high degree might cause a "ping-pong" handover effect. Nevertheless the interference mitigation issues covered in the paper, could be applied to Mobile WiMAX and could be the target of further research[3].

[3]As stated before in this report we will assume that the resulting interference will be handled by network dimensioning and power control and it won't be examined in detail.

All in all load distribution with directed handovers doesn't seem to be a very popular research topic for traditional GSM networks. The main reason for this is probably the fact that resource reservation based load balancing methods, such as channel borrowing, are more feasible because of the centralized and intelligent access network structure. Resource allocation has the advantage over load distribution with handovers, that it is not dependent on MSs residing in the overlapping areas and won't cause handovers since the channel is usually assigned in the beginning of the call. Traffic in traditional networks is homogeneous and therefore load balancing with channel borrowing is simple to manage, whereas in Mobile WiMAX, fluctuating traffic characteristics and distributed access network structure make such resource borrowing very challenging.

Another reason that contributes to the lack of load distribution in traditional networks could be the higher number of non-static MSs. If overlapping areas are small and mobility is high it is quite likely that many MSs won't be spending much time in the overlapping areas. The case, however, is different for WLAN and Mobile WiMAX where more MSs are static (e.g. laptops).

In addition more guard bands (e.g. handover prioritization) and resources in general are reserved for traffic in traditional cellular networks to ensure QoS fulfilment and hence there is not always such a crucial need for load balancing. With the more best effort oriented networks such as WLAN efficient load distribution to maximize the usage of the free resources is more vital. This issue is very interesting with Mobile WiMAX because it will be one of the first systems that comes with an existing support for both differentiated QoS and BE services.

Directed handovers in WLAN and IEEE 802.16e networks

Although in most cases in WLAN the load distribution initiation happens in the MS, there has been some research on Access Point (AP) initiated load distribution. In [Vel04] a directed handover based load balancing scheme for a WLAN AP cluster was proposed. The triggering scheme is quite simple and uses the average (equation (3.1)) and system load balancing index (equation (3.2)) introduced earlier. In the scheme the load balancing index of the system is calculated periodically in each AP. If the index is less than 1, the average load level in the system is calculated and a load state for the AP is computed. The possible load states are underloaded, balanced and overloaded. They are defined as depicted in Figure 3.1.

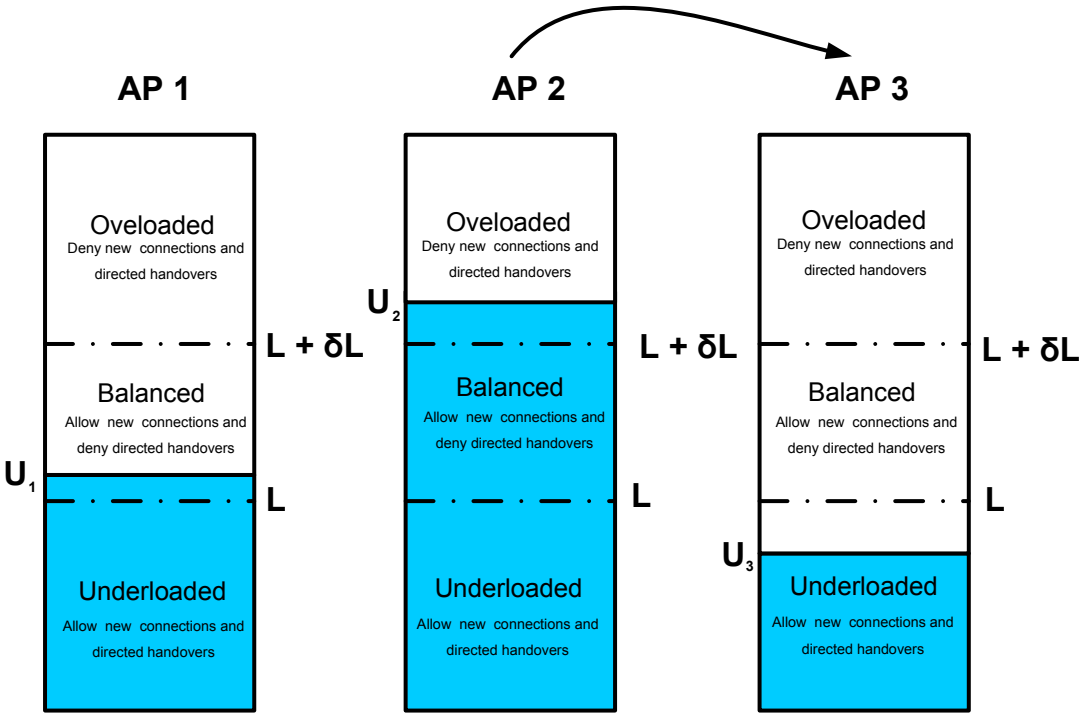

Figure 3.1: Load balancing operation with the scheme in [Vel04].

When resource utilization reaches the overloaded area, load balancing is triggered. The overloaded area is defined as the area passing the threshold $L + \delta L$ where δ characterizes the size of a hysteresis margin. The reason to use such a margin is to combat the effect of unnecessary "ping-pong" handovers due to traffic and channel variations as discussed earlier. The δ parameter defines how much traffic unbalance will be tolerated and can be set in relations to how variable traffic and the channel are.

In the proposed scheme the directed handovers are conducted only from APs that that are overloaded to APs that are underloaded (from AP 2 to AP 3 in Figure 3.1). New service flows are denied in the overloaded state. How often the whole process is repeated, is specified by a Load Balancing Cycle (LBC). The paper also proposes a best candidate approach when choosing which MS to handover. The idea is to handover an MS that is using an amount of resources that would be as close to the difference between the average load L and load of the AP U_i.

Many of the ideas presented in the paper seem feasible also for Mobile WiMAX. What makes it especially attractive is that it is simple to implement and that it takes into consideration the fluctuating characteristic of the resource usage. Couple of adjustments should be however made. In the proposed scheme only one MS per loading cycle is handed over whereas with Mobile WiMAX the number of MSs that

should be handed over could be pre-calculated as discussed in subsection 2.2.2.2. In addition, due to the admission control that will be runnig in the background, new calls could also be admitted in the overloaded BS.

Another interesting directed handover based load balancing algorithm was introduce first in [Moi06a] for WLAN and later in [Moi06b] for IEEE 802.16e based networks. The scheme tries to find the optimal MS-BS association set to balance the utilization of common resources in the whole system. The algorithm goes through every possible association combination, trying to minimize the maximum resource utilization of slot (time and frequency) and power resources in each BS. It has the potential to distribute the load very effectively, but comes with some drawbacks.

The algorithm is a little too complex in the sense that it does not only balance system load but also tries to decrease the load which will increase the number of directed handovers and therefore might endanger QoS fulfillment. The question of *when* directed handovers should be conducted in relations to fluctuating traffic and how much unbalance should be tolerated, is not addressed in this algorithm, as it is in [Vel04].

Also, in order for the scheme to work, information about the possible sets of MCSs and power levels in each association option would have to be acquired and communicated to each decision entity. If the algorithm would be implemented in a distributed manner in the Base Stations (corresponding to profile C), it would cause a high amount of signaling. It would fit in better to a centralized approach (corresponding to profile C) but that would reduce scalability. Many of the ideas presented in the paper, could be used for load balancing in the later stages of Mobile WiMAX deployment but for now, it seems too complex.

Another idea worth mentioning that relates to load balancing with directed handovers in Mobile WiMAX was presented in [Lee07]. There a proposal was made to simplify the handover scanning and network re-entry procedures when conducting directed handovers between the sectors of a cell[4].

All in all there has not been much research conducted in terms of load balancing with directed handovers in Mobile WiMAX so many questions still remain for us to answer. Out of the schemes presented above, the hysteresis based load balancing scheme [Vel04] seems most feasible for our purposes so we will apply that from now on.

3.1.2.2 Directed Retry

What about if all the resources of the BS are used even after load balancing has been conducted? Under these conditions, if an MS residing in an overlapping area is

[4]In the suggested scheme, the MS makes the load balancing decision.

trying to establish a connection and is blocked, it will eventually try to enter another BS. This might however take a long time. The idea behind Directed Retry (DR) is for the BS to explicitly direct the blocked connection to another BS. When the BS assists the MS in the redirection, network entry and connection establishment can be done much faster because similar pre-associations and backbone pre-negotiations can be conducted as with a regular handover. DR can be thought of as a directed handover for a connection that hasn't yet been established.

Although the standard does not support this functionality inherently, it is still a potential enhancement in terms of load balancing and better QoS. Following is a brief review of the previous research conducted in regards to DR.

Directed retry has been researched in the traditional cellular context a great deal. In [Ekl86] directed retry was first introduced. It was proposed that if a user is in an overlapping area, and finds its first-attempt cell has no free channels, it could look for free radio channels in more than one BS as long as the target BS can provide sufficient signal quality.

In [Yum93] it was shown that the use of directed retry, is expected to cause only a minimum amount of additional load in handover processing and has only a minimal effect on the probability of handover failure. [Wat95] further showed that load sharing between the sectors of one cell decreases the blocking rate of new calls as a function of the size of the overlapping area.

The idea of Directed Retry was applied in an interesting way in [Bal02] to future WLAN networks. It introduced the concept of network directed roaming, where the idea is to direct users that are not in an overlapping area and whose connection is blocked to the nearest AP with most free capacity. In other words the BS would give the user co-ordinates where another Access Point is located. This would be an additional way to provide better service to the user and although not currently supported by the standard could also be an interesting feature.

Both directed retry and network directed roaming could be used in Mobile WiMAX with few modifications to the initial network entry procedures.

3.2 Handover and traffic type prioritization

Dropping an ongoing call due to a lack of resources is generally considered worse than blocking a new call. Handover prioritization is used to prioritize existing calls over new ones by reserving resources beforehand for possible rescue handovers. Also different types of traffic can be prioritized with similar guard bands in relations to each other (e.g. UGS based VoIP prioritized over nrtPS based FTP). This gives an operator the opportunity to offer additional QoS in the form of maximum service flow blocking and dropping guarantees (e.g. only 2 % of UGS based VoIP calls

handovers will be dropped in the system). Such measures are already in common use in traditional cellular networks.

WiMAX is the first technology of the 802 family that comes with a designed support for connection oriented traffic enabling true MAC level QoS guarantees. This QoS philosophy could be complemented and enhanced when WiMAX is extended to mobility, by bringing in these system wide connection dropping and blocking QoS guarantees. As it is quite probable that load balancing can only enhance the probability to fulfill these system wide QoS guarantees but not ensure their fulfillment, the usage of the above discussed guard bands should be considered.

In addition, as capacity demands grow, cell sizes will become smaller, which further increases the number of rescue handovers conducted. This makes handover prioritization an even more important issue. In the following subsections we will take a look at both handover prioritization and traffic prioritization.

3.2.1 Rescue handover prioritization

In the following we will first introduce the basic concept of rescue handover prioritization and will then move on to studying some of the previous research conducted.

3.2.1.1 Introduction

The basic principle when admitting new connections is prioritizing the QoS of existing connections over the new connections. In other words we should not admit a new call if it will degrade the level of service received by an existing connection below a certain level.

This is usually ensured by using an admission control scheme which calculates whether there are enough free resources for new calls arriving to a BS. When mobility is introduced, the situation becomes more complex because we have to admit the new connection to the whole system. This means that we have to reserve resources for connections that will experience one or more rescue handovers to other BSs. Handover prioritization will therefore increase the number of blocked new calls and also decrease resource utilization efficiency.

Handover prioritization can be roughly classified to two categories: *Fixed Guard Band Schemes* and *Dynamic Guard Band Schemes*. In Fixed Guard Band Schemes (also known as Cut-off Priority Policy (CPP)) the guard band is fixed and defined in network dimensioning. It can be complemented with throttling, where new calls are randomly blocked (based on a throttling probability), if the rescue handover arrival rate increases, making the scheme a little more adaptable to varying traffic. The advantage of using fixed schemes is that they are simple. As a disadvantage it requires a lot of manual planning and optimization and can also waste a lot of resources when traffic conditions vary.

With Dynamic Guard Channel schemes the idea is to tune the guard channel dynamically based on the number of ongoing calls in neighboring cells, estimation of the channel holding time and the number of handovers to and from the BS. More sophisticated methods can even use mobility prediction for the resource reservation. Often a cluster of neighboring BSs is used in the calculations.

The biggest advantage that the dynamic schemes introduce is more efficient resource utilization without compromising the QoS requirements. Complexity that results from required information exchange between BSs and logic are on the other hand a disadvantage. Signaling can however be reduced by conducting only local measurements of the rescue handovers arriving *to* the BS. Such an approach would also improve scalability and make dynamic schemes more feasible to be deployed in the existing WiMAX Forum network architecture which does not support handover prioritization inherently. All in all the dynamic approach to handover prioritization should fit in better with Mobile WiMAX.

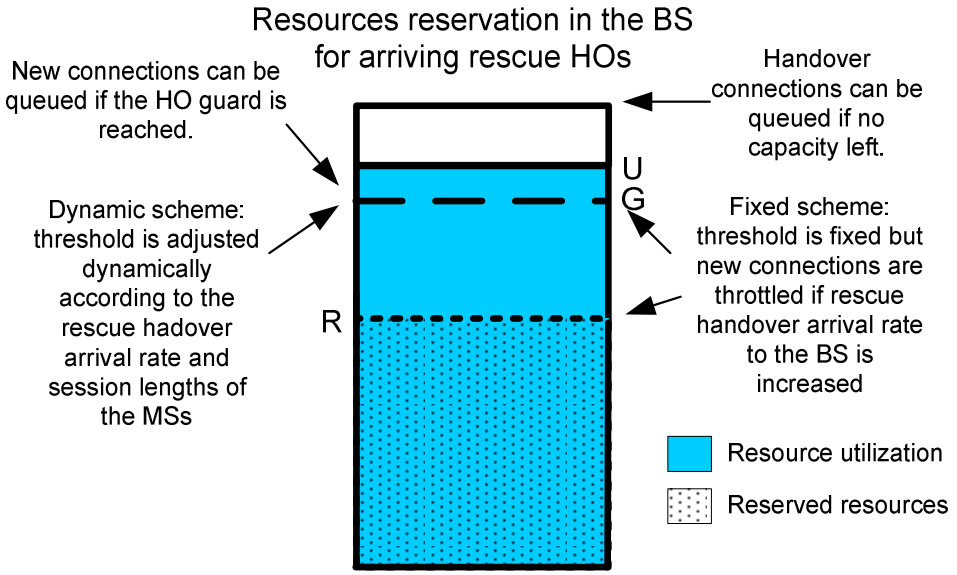

Figure 3.2: Handover prioritization.

The differences between the handover prioritization scheme types are summarized in Figure 3.2. Note that with these guard band schemes, comparisons are made in terms of the *reserved resources R of the BS, not the used resources U* as is done with load balancing. Reserved resources correspond to service flow level arrivals and slot holding times whereas resource utilization corresponds to traffic load on the packet level.

In systems where variable traffic, such as video and Transmission Control Protocol (TCP) based elastic traffic, is served resource utilization can vary a great deal and, as can be seen from the figure, can temporarily pass the guard threshold without causing blocking of connections[5]. At other (e.g. with many VoIP VAD connections) times resource utilization can be less than resource reservation and new call blocking can then occur before resource utilization reaches the guard band. Therefore load balancing can be triggered in relations to resource utilization and resource reservation whichever has the worst case.

So what counts in terms of handover prioritization are the reserved resources. Every service flow (be it from a new connection or a rescue handover) that has acquired a minimum bandwidth reservation guarantee contributes to the total resource reservation. The worst case in resource reservation out of the UL and DL is taken in a similar way as with resource utilization. If the total resource reservation passes the guard threshold, new connections will be blocked.

The new connections can also be queued, after total resource reservation passes the guard threshold. Also, the service flows of the rescue handovers can be queued, if all resources are reserved. Queuing can be done with a traditional First-in-First-out (FIFO) discipline, but also with a prioritized non-preemptive discipline [Xha04] on the basis of traffic priority[6] to prioritize delay sensitive connections or on the basis of the MSs velocity and distance from the Serving BS.

The resource reservation is possible even proactively if the mobility pattern of the MS is known in advance. For example in the case where a Global Positioning System (GPS) device is used to plan the route a car will drive, this information could be utilized for handover resource reservation. Also cognitive radio (802.16m) could introduce a method, where the network learns the trajectories and traffic profiles of the MSs and hence resources could be reserved accordingly. If the mobility patterns are at least partly known and if both macro and micro cells are present in the system, handovers from micro to macro cells can be used to minimize the number of handovers to further mitigate the resource reservation problem.

3.2.1.2 Previous research

Handover prioritization has been a popular research topic and there are many schemes. Here is a brief overview of a few interesting ones for our purposes.

Fixed Guard Band

The idea of reserving bandwidth for handovers was introduced in [Hong86]. In

[5] After the resource needs of the service flows with minimum guarantees (MRTR) are met, all the resources that are left over, can be used by the connections that still have remaining bandwidth left (until MSTR).

[6] E.g. a more jitter sensitive UGS VoIP flow before an rtPS video flow.

[Bar04] most of the existing schemes based on Fixed Guard Band were collected into a State-Dependent Rejection scheme, where the throttling probability (probability that a new call is dropped) can be set differently based on the state the system is in. The state is defined based on how many connections are being served.

Dynamic Guard Band

Roughly speaking it seems that there are two approaches to Dynamic Handover resource reservation. The threshold can be adjusted based on information recorded only locally or based on both local information and information exchanged between adjacent Base Stations. The information exchange based schemes can further be divided into ones that just try to estimate the resources needed in each BS for incoming handovers and ones that enable explicit handover resource reservation per-connection for the entire route that each MS will traverse.

An example of a local scheme can be found from [Lee03] where a simple reactive scheme that adjusts the guard band based on handover-dropping events was presented. This is a very simple scheme, because it does not take mobility into consideration at all. Another locally based scheme was presented in [Cho98] where resources for anticipated handovers are reserved based on estimations made of the rate of arriving rescue handovers. This is a more proactive scheme as it also tries to predict mobility.

There have been many schemes where BSs exchange information to predict the required handover resources. In [Nag95] a Dynamic Guard band scheme, that takes into consideration the number of calls in adjacent cells was introduced. The scheme works in a distributed manner without the involvement of a central network entity. In [Ira01] the question of how many neighboring BSs should provide information to the handover resource reservation was studied. It was concluded that it is definitely worthwhile to involve many adjacent BSs in the decision but the exact number of how many is hard to define. In [Die04] a simple and scalable dynamic handover prioritization scheme for future mobile networks was introduced. In this easily deployable scheme, the guard band is adjusted based on mobility information exchanged between the neighboring cells.

One example of a per-connection reservation scheme was introduced in [Lev97] where the shadow cluster concept used to estimate handover resource reservation, was studied. The idea behind the scheme is to admit only those MSs that are likely to complete their calls in the cluster of BSs. The decision is made based on the QoS requirement and mobility pattern information.

In [Cho00] five dynamic handover prioritization schemes were compared in terms of such measures as dropping probability, new connection blocking probability, bandwidth utilization, and complexity. Three of the schemes were based on per-

connection reservation and two (the above mentioned [Cho98] and [Nag95]) were based on prediction. It was concluded that per-connection bandwidth reservation is too expensive unless it is used in a situation where the mobility patterns are known (e.g. in a highway[7]). Out of the prediction based schemes the local scheme [Cho98] outperformed the information exchange based scheme [Nag95] and was superior especially in terms of complexity and computational demands.

In terms of the existing WiMAX Forum network architecture both approaches, local and information exchange based, are possible but the local approach seems more feasible for Mobile WiMAX Base Stations due to its simplicity and local nature. The WiMAX Forum network architecture does not provide any kind of a framework for information exchange between BSs which means that a new protocol would have to be introduced for an information exchange based dynamic reservation scheme resulting in lower scalability.

A potential area of research that could contribute to enhancing handover resource reservation is mobility pattern estimation. One example was presented in [Liu98] where a hierarchical user mobility model to estimate the mobility pattern of a user was designed.

3.2.2 Prioritizing different types of traffic

As handovers can be prioritized over new calls, traffic prioritization can also be done in terms of different types of traffic. In the following we will have a brief look on the basic idea of traffic prioritization and the previous research conducted.

3.2.2.1 Introduction

Many systems, including Mobile WiMAX, offer different QoS classes enabling connection prioritization. This means that similar guard bands as with handovers can be used in the admission control scheme to prioritize new service flows with higher priority over lower priority service flows (e.g. new UGS VoIP service flow over new nrtPS FTP service flow). Traffic prioritization can be extended also to handover prioritization to further differentiate the service received by the MSs. This means that handovers for a higher priority QoS class would be prioritized over handovers conducted for a lower priority QoS class (e.g. UGS VoIP service flow handover over a nrtPS FTP service flow handover).

Such approach will result in several guard bands based on which the admission control mechanism would block different types of calls. Otherwise the approach is the same as with the single guard band handover prioritization case and the blocking decisions will be made based on the total resource reservation (not resource utilization). In the case where all resource are reserved, the connections can also be queued according to their priority.

[7] Another possibility would be to use it in conjunction with GPS route generation.

The advantage of using such differentiation is that the higher priority schemes can receive even better QoS. It is far more irritating to be dropped out of a conversation than to wait a little longer for your FTP download and therefore the possibility to ensure that handovers for non-real-time connections are dropped before dropping handovers for real-time connections is very beneficial. The slight disadvantage that this approach introduces is additional complexity.

3.2.2.2 Previous research

There has been some research on traffic prioritization. In [Jay00] a framework for QoS provisioning for multimedia services was proposed by using different treatment for real-time and non-real time traffic on the link layer. [Zen00] proposed a handover scheme where priority reservation for voice handover was used. In the scheme resources are reserved for both voice and data handovers but the voice handovers have priority over data handovers.

In [Xha04] a framework for dynamic priority queuing was presented. Although in the paper, priority was based on the received signal strength and the remaining time in the overlapping region between two cells, it could potentially be used for prioritizing also different types of traffic.

In [Che05] the idea of prioritizing handovers based on their traffic type presented in [Zen00] was applied to a dynamic environment with a dynamic multiple-threshold bandwidth reservation (DMTBR) scheme. The scheme uses a dynamic guard band for handovers while maintaining relative priorities for different traffic classes. It is capable of granting differential priorities not only to different traffic classes but also to new and handover traffic for each class by dynamically adjusting three bandwidth reservation thresholds. The scheme assumes two traffic classes non-real-time (nrt) and real-time (rt). The thresholds used in the scheme are presented in Figure 3.3.

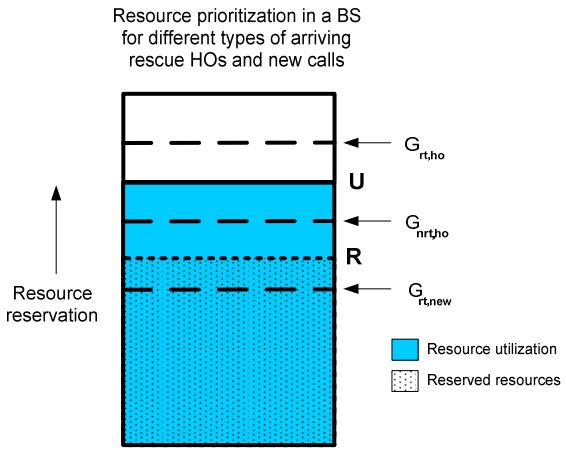

Figure 3.3: Multiple-threshold bandwidth reservation [Che05].

As resource reservation increases, the resources reserved after the guard band for new real time connections has been passed, can be used by new rt connections[8] and by nrt and rt handovers. In the same way the resources reserved after the guard band for non-real-time handovers can be only used by nrt and rt handovers. Finally, the resources reserved after the guard band for real-time handovers can only be used by rt real-time handovers. All new nrt connections will be blocked after the new real-time connection guard band has been passed which will happen in the example in Figure 3.3.

The proposed scheme works locally by first estimating initial values for the thresholds based on instantaneous mobility and traffic load situation. The thresholds are further adapted according to instantaneous QoS measures such as dropped handovers and blocked new calls in a similar way as in [Lee03]. Throttling (blocking new calls randomly) is also used when network becomes congested.

[8]The reserved resources could be used e.g. for the MCS changes for the rt connections.

Chapter 4

Load Balancing with Handovers in Mobile WiMAX

Now that we have a good understanding of the background behind load balancing with handovers and behind rescue handover and traffic prioritization and have also covered the key system aspects of Mobile WiMAX in relations to these, we can start to consider how load balancing and rescue handover and traffic prioritization could actually be applied to Mobile WiMAX. In this chapter, we will first design a basic load balancing scheme by applying an existing scheme from prior research to Mobile WiMAX and see how it could be enhanced in the Mobile WiMAX system. Second we will examine how the basic algorithm could be further complemented with handover and traffic prioritization.

4.1 A Load balancing algorithm for Mobile WiMAX

We will begin by applying the algorithm presented in [Vel04] to Mobile WiMAX. The scheme was originally introduced for WLAN networks so it will need some modifications and adjustments from our part in order to be compatible with Mobile WiMAX.

4.1.1 Assumptions for the algorithm

As already discussed in the previous sections the algorithm will run in a distributed manner on each BS corresponding to ASN profile C. The aim of the algorithm is to use directed handovers to balance the load offered to the system thus improving global Resource Utilization and decreasing the number of blocked new service flows and rescue handover drops. The aim is also to tolerate some unbalance in the system to reduce unnecessary directed handovers.

As mentioned before, we will assume that interference is not a critical issue for the MSs in overlapping cells, and that it will be dealt by proper power control and

radio network planning[1]. Also we will assume that in the basic algorithm the hysteresis margin is set manually[2].

The metric used with the algorithm will be slots as defined in the WiMAX Forum network architecture [ASN2]. Resource Utilization will be measured and reported only for the non-BE service flows and hence BS initiated load balancing handovers will also be conducted only for MSs with non-BE service flows. Static DL and UL subframe division will be used and hence the more loaded subframe will determine Resource Utilization as defined in equation (3.3)[3].

In addition it might be beneficial to conduct load balancing only for MSs that are likely to reside in the overlapping areas for their whole session. MSs that move with a high velocity, are likely to move between cells and conduct rescue handovers during their session, which might result in unnecessary handovers if directed handovers are also used. Rescue handovers, especially with high velocities, are challenging to perform because they usually require heavy execution mechanism such as FBSS or MDHO. We will therefore conduct load balancing only for static MSs[4]. The identification of whether a new MSs is static or mobile, could be based on round trip delay measurements, variations in the channel measurements, or on mobility prediction.

4.1.2 Description of the load balancing algorithm for Mobile WiMAX

The algorithm will mostly work as described originally in [Vel04] but a couple of changes will be made. First of all, different from the original scheme, we will make a distinction between rescue handovers and directed handovers so that they can be treated differently by the TBS. Rescue handovers will be admitted in all loading states but directed handovers only in the underloaded state.

Secondly, in the original scheme new calls were rejected in the overloaded state but since in Mobile WiMAX an admission control scheme will work in the background to ensure that the minimum guarantees for the existing connections are still fulfilled, also new calls can be admitted in the overloaded state. The load balancing index β won't be used at all since all users are not in overlapping areas and it will be quite unlikely that the system will be totally balanced. The average load of the system and the loading states will be therefore computed each Load Balancing Cycle (LBC). The basic framework of the algorithm is described in Figure 4.1.

[1]This issue is outside the scope of this thesis, but could be addressed in the future.

[2]Later in part 4.1.3.1 we will address automatic tuning in more detail.

[3]If dynamic DL and UL subframe division is used the calculations can be done for the total number of slots.

[4]Or at maximum to MSs that move with a low velocity.

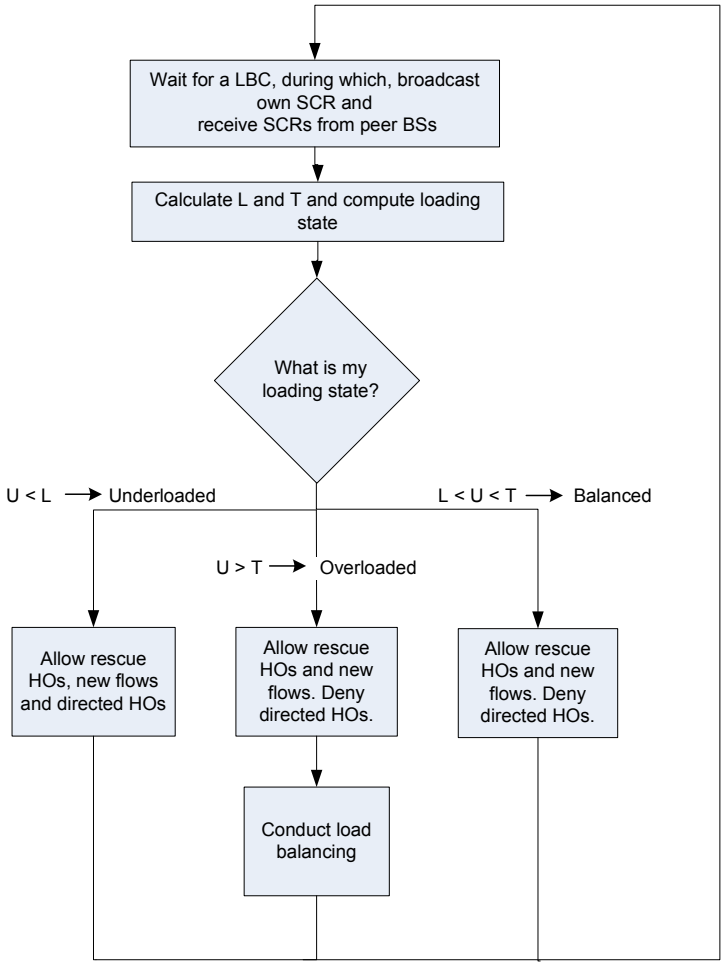

Figure 4.1: Overall logic for the basic load balancing algorithm [Vel04].

A Spare Capacity Report will be broadcasted every LBC. The length of the LBC can also be used as the averaging time for the SCR Resource Utilization measures described in part 2.2.2.2. During the LBC the BS receives SCRs from its peer BSs, computes loading states for the peer BSs and measures its own Resource Utilization. Here, it might be beneficial to wait as long as possible to send the SCR, so that the most up to date information is reported.

After the LBC has ended the average load of the system L and the threshold to trigger load balancing $T = L + \delta L$ is calculated. Based on this the loading state of the BS is computed, by comparing the average Resource Utilization U to the average load L and to the triggering threshold T[5].

[5]Since all measurements are reported in percentages, comparisons between BSs with different capacities can be made.

For the next LBC, incoming requests for new service flows and rescue and directed handovers will be treated based on the loading state as described in Figure 4.1. If the BS is in an overloaded state load balancing will be triggered[6].

When load balancing is triggered the BS will initiate directed handovers for MSs that reside in overlapping areas between Base Stations. What the BS needs to decide is *which* MSs, in *what order* and *how many at a time* will be handed over. As discussed in 2.1.3.3 the length of the load balancing procedure depends on whether the BS is already aware of which MSs are in the overlapping areas. The logic that could be used during load balancing is presented in the sub block diagram in Figure 4.2.

After initiating load balancing the BS will have to find out *which* MSs are static and in an overlapping area. If no ready list exists of these MSs they have to be discovered before directed handovers can be initiated. To reduce unnecessary scanning the first step could be to narrow down the candidate MSs to ones that are static and likely to reside in an overlapping area. This could be done by using measurements on channel variation, signal strength, round trip delay and also by using location estimation methods[7] [Liu98][Bah00].

A cell re-selection procedure could be initiated for the remaining MSs by sending them unsolicited MOB_SCN-RSP messages telling them to scan all neighbor BSs based on the info received in the MOB_NBR-ADV message. The results could be reported via the radio interface from the MS to the SBS, or with the Physical Parameters report from the TBS to the SBS. Based on the results a list of MSs that are in an overlapping area (within the signal range of at least two BSs) will be generated. Also the set of TBSs with feasible signal strengths will be recorded for each MS. If a list of overlapping static MSs would be kept before load balancing is triggered it could be based on a similar procedure.

After the list of static MSs in the overlapping area is ready, the list could be further pruned and the MSs in the list could be prioritized. For example the MSs that have candidate TBS sets where none of the TBSs are in an underloaded state can be removed.

[6]Conducting load balancing might not be beneficial when the system is extremely loaded or only lightly loaded so an additional $L_{min} < L < L_{max}$ check could be made.

[7]Location estimation could theoretically be used alone, but since it is still too inaccurate, it should only be used as a complementing method.

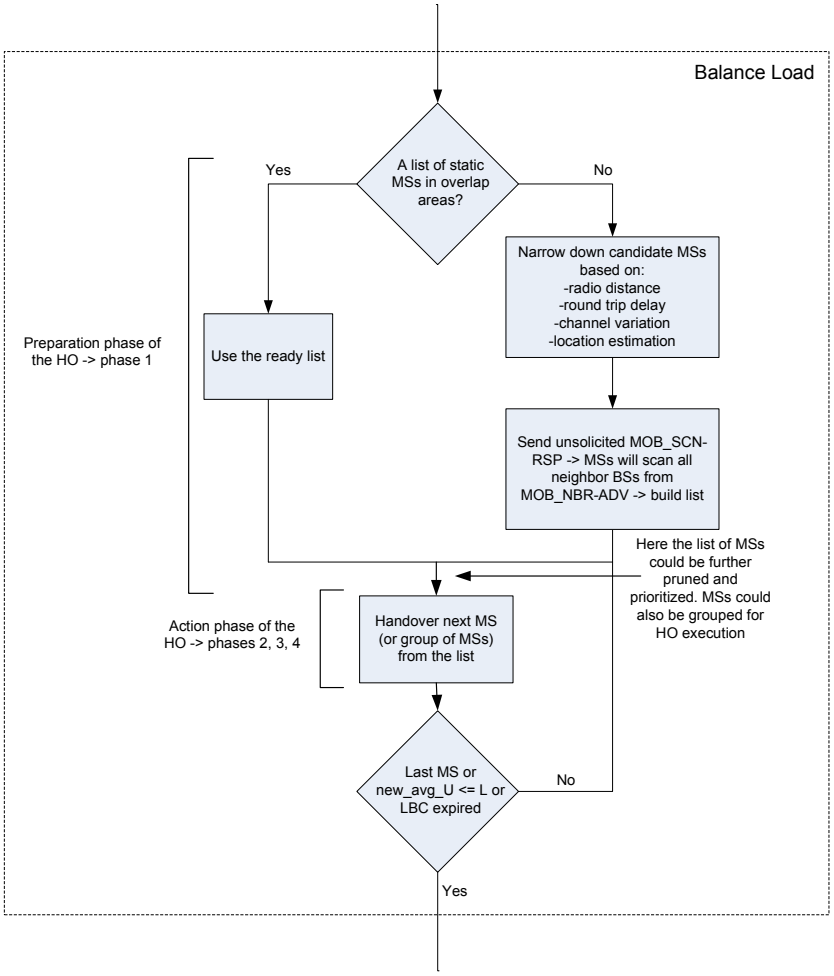

Figure 4.2: Logic for the basic load balancing algorithm when directed handovers are triggered.

When conducting directed handovers, the TBSs might eventually go to the balanced state and will start to deny incoming directed handovers. Therefore the most critical MSs whose directed handover cause least disturbance could be handed over first. In traditional networks, where traffic is rather static and overload situations clear, the higher priority connections have usually been handed over first. However when the traffic starts to be very fluctuating it might actually be beneficial to hand over the most delay sensitive connections (e.g. VoIP) last, to avoid unnecessary "ping-pong" handovers as long as proper admission control and scheduling schemes that enforce the prioritization of connections in the congested BS are working in the background. We will study this in further detail later in 4.1.3.3.

The MSs could also be prioritized based on their radio distance, Physical Service Level in the TBS or resulting interference [Fuj92]. The best candidate approach dis-

cussed in the original scheme [Vel04], as well as the per QoS profile Spare Capacity Reporting procedure that will be introduced in the later stages of Mobile WiMAX could also be used here for decision support.

After prioritization of MSs has been done the MSs could be grouped so that handovers could be executed in parallel. Using such groups would reduce the time used for load balancing but would run the risk of collisions in the network re-entry procedures if the groups were too large. The prioritization and grouping discussed above is an enhancement to the basic scheme. In the basic scheme we will not prioritize the MSs and will simply handover one MS per frame.

Everything up until now has been part of the handover preparation phase as defined in the WiMAX Forum network architecture (cell reselection phase in terms of the radio link) and as the directed handovers are initiated we will move to the handover action phase (initiation and execution phases in terms of the radio link). The BS will initiate the directed handover by sending a HO_req message to the candidate TBSs and the procedure will proceed as described in Figure 2.9. The TBS will deny or admit the directed handover based on its loading state and inform about it in the HO_rsp message. If there are many TBSs remaining, the MS will make the final decision where it wants to handover to and can perform additional scanning and associations if necessary before sending the MOB_MSHO-IND message.

As discussed before to make the load balancing locig work, it would be necessary to specify in the HO_req message, whether the handover in question is a rescue or a directed handover. Currently the HO type field in the HO_req message only indicates the handover type in terms of either hard, FBSS or MDHO handover. The remaining bits could be used to differentiate between a BS initiated directed handover and a MS initiated rescue handover.

The next MS (or a group of MSs) from the list will be handed over until the end of the list has been reached, the new resulting Resource Utilization new_avg_U is equal or below the average L, or the end of the Load Balancing Cycle has been reached. The new resulting Resource Utilization new_avg_U can be calculated using the average Resource Utilization of the released service flow. The reason that the current Resource Utilization measurement is not used is that the effect of the released resources won't be shown immediately in the measurements because they are averaged. A similar problem can occur in the TBS, where the new service flow will be created. To reduce unnecessary handovers, an estimation of the average Resource Utilization of the new flow can be added to the measured average Resource Utilization.

4.1.3 Possible enhancements

Before moving on to the handover and traffic prioritization enhancement part we will take a look at some enhancements that could be made to the basic algorithm

in terms of automatic computation of the triggering threshold, BS initiated load balancing for BE MSs and how multiple triggering thresholds could be used to address negative effects of fluctuating traffic.

4.1.3.1 Automatic tuning of the triggering threshold

In the basic algorithm the hysteresis margin will be set manually and no method to automatically set the load balancing threshold was given. Here we will propose a preliminary framework on how to dynamically adjust the triggering threshold based on the current traffic characteristics of the system. The challenge when setting the threshold in relations to Resource Utilization is on the other hand to avoid unnecessary (ping-pong) handovers resulting from a low threshold and premature reaction to variable traffic, but on the other hand to avoid long delays and packet drops by the BS that occur if the threshold is large and load balancing is triggered too late.

As mentioned in part 2.2.2.2, the Spare Capacity Report includes a Radio Resource Fluctuation value F that describes the degree of fluctuation in channel data traffic throughputs for the Base Station. This value ranges from a minimum 0 corresponding to traffic mix of UGS based VoIP connections with steady channel conditions to a maximum 255 corresponding to a traffic mix of highly varying traffic sources with varying channel conditions. In other words the more mobile the served terminals are and the more variable traffic[8] they have, the higher value will be reported.

As a basis to automatically compute the triggering threshold two boundary values $T_{U,min}$ and $T_{U,max}$ could be set. The lower boundary value $T_{U,min}$ includes a minimum hysteresis margin required to avoid the ping-pong effect resulting from one BS initiating and another accepting too many load balancing handovers (we will call this the handover based ping-pong effect). Note that this ping-pong effect caused by the MSs being handed over is different from the ping-pong effect caused by general Resource Utilization fluctuation[9] (we will call this the fluctuation based ping-pong effect). The former is caused by incorrect estimates of the number and Resource Utilization of the MSs that are handed over and accepted and the latter by all traffic and channel fluctuation in the BSs.

$T_{U,min}$ could be set in relations to the average system load L and average system Radio Resource Fluctuation F_{sys}, and will increase as F_{sys} increases. F_{sys} could be calculated based on the values received from the SCR of other Base Stations thus describing the overall fluctuating nature of the incoming traffic.

[8]Roughly speaking as an example we can say that traffic fluctuation increases from UGS based VoIP, to ertPS based VoIP with VAD, to rtPS based streaming video, to nrtPS based elastic FTP and Hyper Text Transfer Protocol (HTTP) traffic.

[9]Similar ping-pong effect can also be seen in the signal based handover decision and there such an unnecessary handover is defined as a situation where the previous link (BS) would have continued to give satisfactory performance [Mar99].

The upper bound reference value $T_{U,max}$ is based on the reliability and performance of the scheduler and denotes the maximum value for the triggering threshold after which the service of the existing connections starts to degrade.

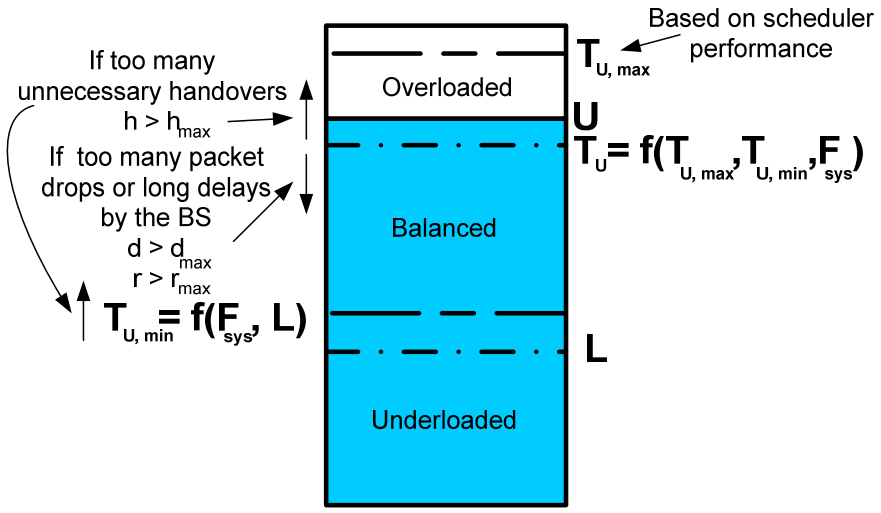

Figure 4.3: Automatic triggering threshold tuning.

So an estimation of the new Resource Utilization threshold can be computed every Load Balancing Cycle as a function of the above mentioned variables: $T_U = f(T_{U,min}, T_{U,max}, F_{sys})$. One example of a simple way to compute the threshold would be with the following equation

$$T_U = T_{U,min} + (T_{U,max} - T_{U,min}) \frac{F_{sys}}{F_{max}} \tag{4.1}$$

where F_{max} is the maximum fluctuation value 255. As can be seen, as the system fluctuation F_{sys} increases the size of the hysteresis margin increases so that the system won't react prematurely to the varying traffic. Both the lower boundary value $T_{U,min}$ and resulting threshold T_U can be reactively tuned in relations to a maximum value for the number of handovers per MS[10] (h_{max}) as depicted in Figure 4.3. The resulting threshold can also be tuned in relations to maximum values for the number of dropped packets (r_{max}) and overlong delays[11] (d_{max}).

The increase of fluctuation in Resource Utilization can also be relieved by increasing the averaging interval used to measure the Resource Utilization. However this has to be done with care as it might make the system too slow to react to varying traffic.

[10] A value defining the maximum number of handovers per minute could be used.

[11] A more specific method that computes and reactively tunes the boundary values $T_{U,min}$ and $T_{U,max}$ and computes and further tunes the triggering threshold T_U could be the target of future research.

4.1.3.2 BS initiated load balancing for BE users

In the WiMAX Forum network architecture, BS controlled load balancing is apparently conducted only for MSs using non-BE services meaning that MSs with only BE service flows are responsible for conducting load balancing themselves. Although the specification does not support the reporting of Resource Utilization of BE users, it could be implemented separately by a BS vendor[12]. The basic algorithm described in section 4.1 could then be applied for the BE users in the following way.

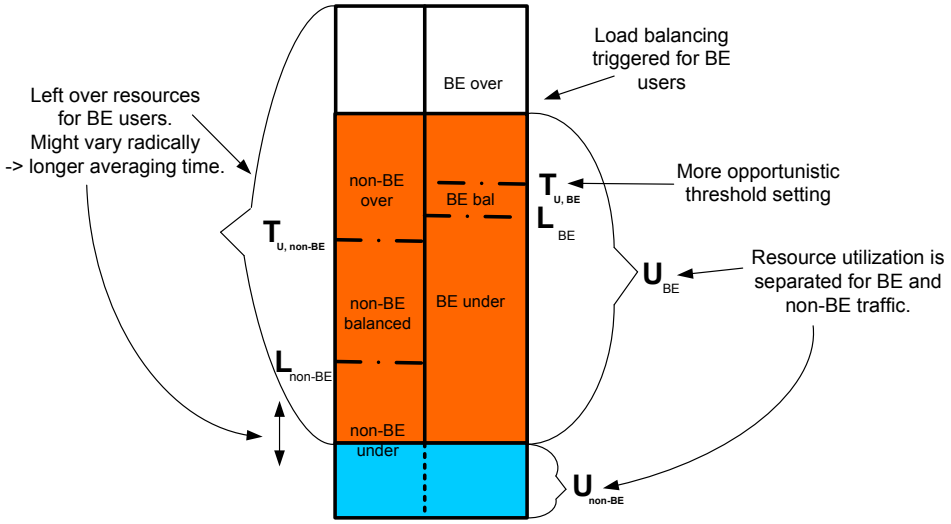

Figure 4.4: BS initiated load balancing for BE MSs in Mobile WiMAX.

Since resources are first utilized by non-BE users, BE users will use whatever is left. This means that the available resources for BE users varies. Same loading states could be computed for the BS in terms of BE users if loading information (free resources and used resources or the MSTRs of the users) of BE users was communicated between the BSs[13]. If another BS has a large amount of resources available for BE users (in BE underloaded state), some of the BE users could be handed over to that BS[14].

The amount of resources available for BE traffic depends on the Resource Utilization of non-BE users and therefore the capacity that the BE connections get might vary considerably. Also the fact that BE traffic is often very fluctuating further increases variability in estimating the loading information. Hence it might be beneficial to use a little longer averaging time (and Load Balancing Cycle) to measure the BE

[12] Additional fields could be added to the Spare Capacity Report.

[13] Not currently supported by the WiMAX Forum network architecture.

[14] Prioritization of which of the BE connections will be handed over could be made based on the MSTR provisioned for the user.

Resource Utilization and resources available for BE users. Still the averaging time should be such that the system is able to react quickly to changes.

Since handovers aren't such a critical issue for the BE MSs, ping-pong handovers could actually be utilized to get access to more bandwidth and therefore the triggering threshold could be set in a more opportunistic way than with non-BE connections. The hysteresis margin for BE MSs could be smaller, so that load balancing would be triggered earlier and the BE users would able to benefit from the BSs that have most capacity left for BE users[15].

The tuning method introduced above in part 4.1.3.1 could also be used here and the smaller, more opportunistic hysteresis margin could be set by choosing a lower upper boundary reference value.

4.1.3.3 Multiple threshold triggering in a fluctuating environment

As already stated unnecessary ping-pong handovers that result from premature reaction to fluctuating radio resources pose a great threat to the QoS of delay sensitive connections such as VoIP which are sensitive to scanning and require heavy handover mechanisms. The simple solution where the averaging period is just increased, will make the system slow to react to traffic variations and decrease system wide Resource Utilization.

Although traditionally, with rather static traffic conditions, the higher priority connections have been handed over first to the less congested cell, in a fluctuating environment it might actually be beneficial to handover the delay sensitive connections last. This way the delay sensitive connections avoid unnecessary handovers and the delay tolerant connections have a chance to react to the load increase and get higher bandwidth from a less congested BS.

Therefore it would be beneficial if load balancing would be triggered gradually, as Resource Utilization increases, first for most delay tolerant connections (e.g. nrtPS based FTP) and last for most delay sensitive connections (e.g. UGS based VoIP). Traffic prioritization within the classes could be still used so that for example a higher priority nrtPS FTP connection would be handed over before a lower priority nrtPS FTP connection, so that it would have access to more bandwidth.

We will use the automatic tuning scheme presented above in part 4.1.3.1 as a basis, and will use two traffic classes, real-time (rt) and non-real-time (nrt), to present our triggering scheme. To make the rt connections most robust against traffic fluctuation we will set the load balancing triggering threshold for rt to be the same as calculated in the basic scheme $T_{U,rt} = T_U$. The threshold for the nrt class will be

[15]Since most of BE traffic is client-server type (FTP, HTTP), it might be a good idea to make the opportunistic decisions based on DL Resource Utilization (just as long as we have enough UL capacity for acknowledgments).

set to a lower value. One simple way of computing such a threshold would be with the following equation[16]

$$T_{U,nrt} = T_{U,min} + (T_U - T_{U,min})\frac{h_{sen}}{h_{nrt}} \tag{4.2}$$

where h_{nrt} corresponds to the maximum number of handovers allowed per minute for the nrt class and h_{sen} indicates the maximum number of handovers allowed for the most delay and handover sensitive class which in this case would be h_{rt}[17]. An example of the scheme is presented in Figure 4.5.

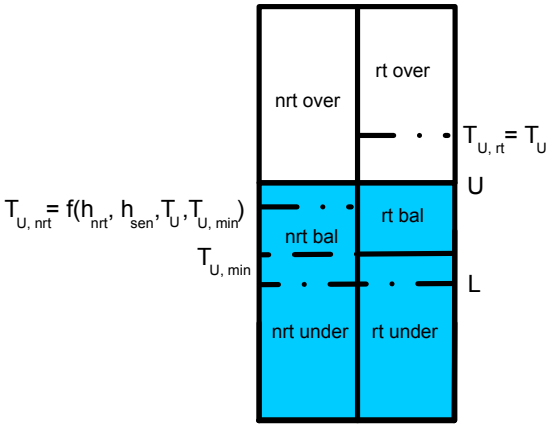

Figure 4.5: Multiple threshold triggering.

In this example load balancing would be initiated for the nrt class and directed handovers would be conducted to the TBSs which would be in the nrt underloaded state[18]. In a similar way, all the other state rules used in the basic scheme (see Figure 4.1) are used here per traffic class.

If the load increase would be only temporary the delay and handover sensitive rt connections would be spared from an unnecessary handover. Furthermore if after a period of time, one of the TBSs load would temporarily increase, the nrt connections would be handed over back to the original cell. This "visit" would be beneficial to the nrt connections because they had access to a larger amount of bandwidth than what they would have had in the original BS.

[16] A more accurate method to compute the multiple thresholds could be the target of future research.

[17] In Mobile WiMAX we could assume that $h_{nrtPS} > h_{rtPS} > h_{ertPS} = h_{UGS}$.

[18] The threshold $T_{U,nrt}$ could be reactively tuned with maximum values h_{nrt}, d_{nrt} and r_{nrt}.

The disadvantage of using this scheme is that when directed handovers for the more delay and handover sensitive connections are finally conducted some of them might be denied since the TBS could already be in the balanced state[19]. However assuming that admission control is working properly, the QoS for the existing connections shouldn't easily degrade even though the BS gets congested. If for some reason (e.g. degrading channels) too many connections have been admitted to the system, the QoS of the other more delay and handover tolerant classes will degrade first (nrt before rt), making them more critical to be handed over to the less congested cell.

An additional advantage of using the multiple threshold approach is that since the UGS and ertPS connections usually reserve and use less bandwidth than rtPS and nrtPS connections, handing over rtPS or nrtPS connections releases more resources in the congested BS. Furthermore the UGS and ertPS based flows require only a certain guaranteed rate and don't benefit from the extra bandwidth available in a less congested BS as much as rtPS and nrtPS connections do.

In addition, with the multiple threshold approach ready lists of MSs in overlapping areas could be kept only of the delay tolerant MSs not sensitive to scanning thus minimizing scanning for the delay sensitive connections. Also arriving rescue handovers that need a heavy execution mechanism and the fact that handovers in general can, at least in the early stages of Mobile WiMAX, be somewhat unreliable, contribute to the reasoning that handovers should be minimized for the delay and jitter sensitive flows.

As stated before the scheme is efficient in a fluctuating environment, but if traffic is rather static and the load difference between the BSs is clear (not a great chance for unnecessary ping-pong handovers) a single threshold scheme where delay sensitive connections are handed over first, should be used[20].

4.2 Complementing the load balancing algorithm with guard bands

Load balancing will improve the possibility to fulfill QoS requirements but cannot itself guarantee anything. Therefore the use of schemes that prioritize different kinds of traffic in terms of mobility and traffic type are most likely needed. In this section we will examine how the basic load balancing algorithm introduced above could be complemented by handover and traffic prioritization.

We will first consider what kind of a handover prioritization method should be used in Mobile WiMAX, create a framework for triggering load balancing in relations to

[19] In this scheme both rt and nrt have the same underloaded state but the problem could be addressed by setting a lower underloaded state threshold for the nrt class.

[20] A threshold in system wide radio resource fluctuation could be set to trigger the multiple threshold scheme.

the handover guard band and examine how directed retry could be used after all possible load balancing handovers have been conducted. Then we will continue on by extending this enhanced load balancing framework to traffic type prioritization based provisioning where multiple guard bands are used.

4.2.1 Handover prioritization and load balancing

4.2.1.1 Handover prioritization in Mobile WiMAX

We will first do some general considerations on what kind of a handover prioritization scheme could be used in Mobile WiMAX. We won't design any detailed scheme, but just present the basic framework for it.

As discussed in part 3.2.1 the handover prioritization scheme in Mobile WiMAX should be distributed and local[21], so that it complies with the WiMAX Forum network architecture and enhances scalability. Dedicated Resource Reservation in the next cell, let alone in an entire shadow cluster, is expensive and doesn't fit well to Mobile WiMAX at least in the early stages.

Due to the flexible nature of Mobile WiMAX, dynamic guard band adaptation based on mobility[22] and traffic intensity in the neighboring BSs[23] is a natural choice as a basis for handover prioritization. Since efficient Resource Utilization is a crucial issue in Mobile WiMAX we don't want the guard band to be too conservative. Therefore a scheme that uses some kind of an initial prediction for the guard band and then reactively adapts it, based on how QoS guarantees, such as handover dropping rate, are fulfilled could be good for Mobile WiMAX. Such an approach would also be very simple.

What is especially interesting to us, in terms of load balancing, is how large the guard band is and how much it will vary, since load balancing can also be triggered in relations to the guard band.

4.2.1.2 Triggering load balancing in relations to the handover guard band

In the schemes discussed above load balancing is triggered in relations to high Resource Utilization (Case 1 in Figure 4.6). It is however possible, when new flows are initiated with a rapid rate, that all resources become reserved before it is shown in the Resource Utilization measurements. This is bad because load balancing won't be triggered to release resources and admission control will unnecessarily start to block calls.

[21] No BS to BS signaling except for the Spare Capacity Report.

[22] Rescue handovers conducted *to* the BS.

[23] Although the SCR won't give any information on the reserved resources, the Resource Utilization can be still used as a general indicator.

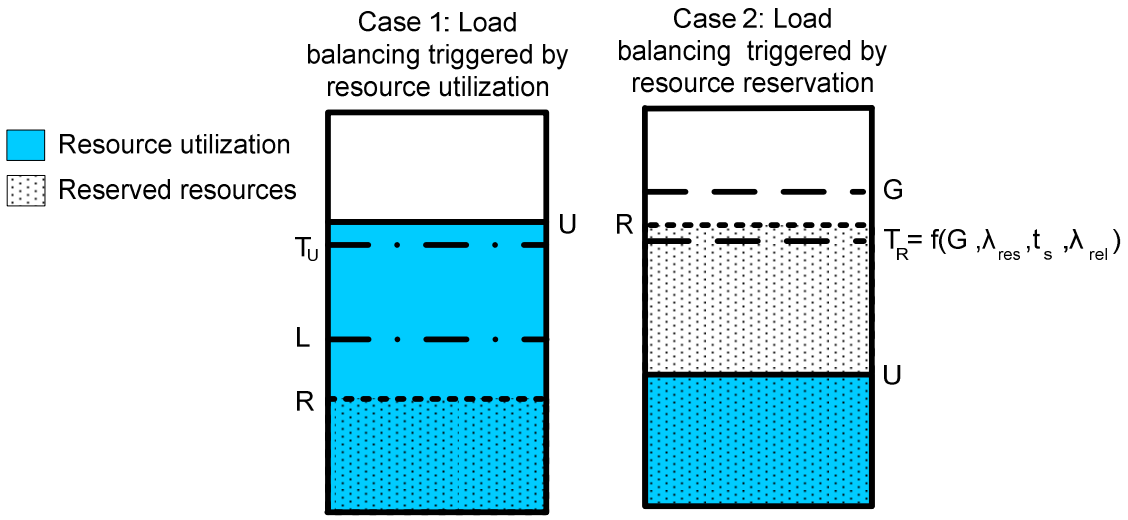

Figure 4.6: Resource Utilization and Resource Reservation based triggering.

Hence it will be beneficial to be able to trigger load balancing also in relations to the guard band for handovers[24] (Case 2 in Figure 4.6). An example method to set the Resource Reservation based triggering threshold is discussed next. If λ_{res} is the average arrival rate of the new slot reservations and t_s is the average holding time of a slot, we can use Little's formula to calculate the number of reserved slots when the system is balanced

$$N = \lambda_{res} t_s \tag{4.3}$$

We can use this to compute an estimation of a threshold for triggering load balancing in relations to current Resource Reservation

$$T_{R,est} = G - (N - G)\frac{\lambda_{res}}{\lambda_{rel}} \tag{4.4}$$

where λ_{rel} indicates the rate at which the load balancing scheme can release slots. As can be seen the higher N and the lower λ_{rel} are the earlier load balancing will be triggered. Since measurements can be inaccurate, we want to trigger load balancing at latest when Resource Reservation reaches G and hence the final triggering threshold will be

$$T_R = \min(T_{R,est}, G) \tag{4.5}$$

So, the idea is to trigger load balancing before G is reached, but not too early to avoid unnecessary handovers[25]. The value of λ_{rel} depends on the duration of the discovery process used to identify overlapping cells and the handover mechanisms used. Since the handover guard band G might also vary, threshold setting can be a challenging task. The threshold could be further reactively tuned in relations to

[24] If no handover guard bands are used load balancing can be triggered in relations to maximum capacity or other possible guard bands (e.g. reserved for MCS changes or MAC headers).

[25] Some fluctuation can happen also on the flow arrival level.

a maximum call blocking rate value b_{max} indicating the case where handovers were triggered too late and unnecessary handover rate value h_{max} indicating when handovers were triggered too early.

When load balancing is triggered based on Resource Reservation the logic from the basic load balancing algorithm will not apply since it is based on Resource Utilization. Deciding which and how many MSs to handover and to which TBS is tricky with the current network architecture since not much Resource Reservation information is communicated between the BSs[26]. The per QoS profile Spare Capacity Report could be utilized to some degree to determine which MSs to handover.

In any case directed handovers could be initiated with HO_req messages and the admission control of the TBS could respond according to its Resource Reservation situation. The handover type could be differentiated from the regular (Resource Utilization based) directed handover by using the additional bits in the HO_req handover type field. The arriving Resource Reservation based directed handovers could be treated as new calls in the TBS up until a certain point[27]. This way the flow arrival burden experienced by one BS would be distributed to the other BSs of the system.

4.2.1.3 Network directed retry and roaming

What about when increasing Resource Reservation has triggered load balancing and all possible load balancing directed handovers have been conducted? As discussed in part 3.1.2.2, in such a case, network directed retry and network directed roaming are potential methods to balance the load of incoming new flows. The basic idea is presented in Figure 4.7.

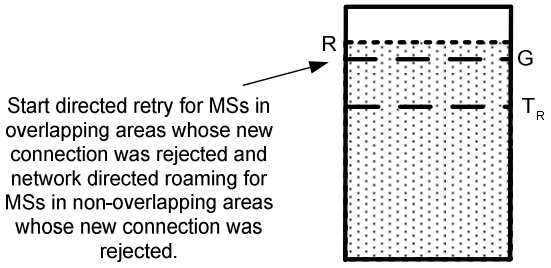

Figure 4.7: Network directed retry and network directed roaming.

[26] Again if a single Resource Reservation based triggering threshold is used, if the load difference between the BSs is clear and if Resource Reservation does not fluctuate that much, the highest priority connections (and most delay sensitive) should be handed over first to the less congested cell.

[27] A hysteresis approach could be used here also.

To make directed retry and network directed roaming work in Mobile WiMAX a few modifications to the initial network entry procedures should be made. When blocking occurs in a BS, a DSA_RSP message could be sent to the MS initiating the service flow with an indication that directed retry or network directed roaming could be conducted. After that a similar discovery process to find out if the MS is in an overlapping area as desribed in the basic algorithm (see Figure 4.2), could be carried out resulting in a directed handover if the MS is residing in an overlapping area. Network directed roaming would be conducted as a last resort for the MS that is not in the overlapping area by communicating the location of the closest lightly loaded BS[28]. This would however require co-operation with application level protocols.

4.2.2 Traffic prioritization and load balancing

Finally we will take a look at what kind of a relationship traffic prioritization has with load balancing. As was examined in part 3.2.2 guard bands can also be reserved to prioritize traffic classes in relations to each other. If such prioritization is used we can use the framework created earlier in part 4.2.1.2 to trigger load balancing in relations to the guard bands.

Here we will complement our load balancing approach with the early discussed multiple guard band scheme presented in part 3.2.2.2 [Che05], where a mixed real-time (rt) and non-real-time (nrt) provisioning method was considered.

In the scheme crossing the threshold protecting new rt connections will cause new nrt connection blocking and crossing the threshold protecting rescue handover nrt connections will cause blocking of new rt connections. By applying the equation (4.5) to these two, we can determine two Resource Reservation thresholds $T_{R,nrt}$ and $T_{R,rt}$, which define when Resource Reservation based load balancing should be triggered for both classes[29].

As Resource Reservation increases load balancing directed handovers are conducted first to the nrt class, reducing the number of handovers conducted by the higher priority, more delay and handover sensitive rt connections (e.g. only a temporary peak in the flow arrivals). Since Resource Reservation based directed nrt handovers will be treated as new calls in the less congested TBSs they cannot use the resources reserved for rt connections.

[28] Could be included in the DSA_RSP or MOB_NBR-ADV message.

[29] The slot release rate will be different for the classes ($\lambda_{rel,rt}$ and $\lambda_{rel,nrt}$) since different handover mechanisms are used.

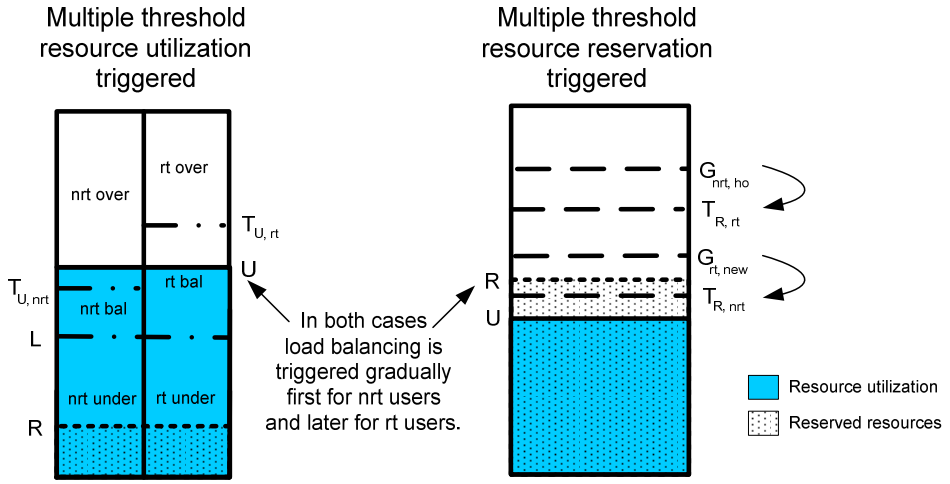

Figure 4.8: A Resource Utilization and reservation based multiple threshold triggering scheme.

As can be seen from Figure 4.8 this Resource Reservation based multiple threshold triggering scheme has many similarities with the Resource Utilization based multiple threshold triggering scheme presented in part 4.1.3.3. In both cases load balancing is triggered gradually minimizing the handovers of most delay and handover sensitive connections while still enhancing system wide Resource Utilization. Although the examples for both of the schemes have been presented with the nrt and rt example classes, these schemes can be easily extended to all of the scheduling services that Mobile WiMAX supports by adding more thresholds.

If handover and traffic guard bands would be used in the Mobile WiMAX system and if the radio resource usage in the system would fluctuate a great deal, the load balancing scheme used could be a combination of these two, reacting to the loading situation on the level[30] that is at the time most critical.

[30]Packet level Resource Utilization vs. flow level Resource Reservation.

Chapter 5

Evaluation

In this chapter we will conduct preliminary evaluation of the basic load balancing algorithm for Mobile WiMAX presented in 4.1.2 in a static environment with a mix of rather static VoIP based non-BE traffic and FTP and HTTP based BE traffic[1]. In the first section we will introduce ourselves to the NS2 based simulator we will be using in terms of its features and configuration. Then we will move on to describe the topology and traffic setup of the simulation scenarios and finally will take a closer look at the actual simulation scenarios and the performance indicators used.

In the second part we will present and analyze the simulation results of each evaluation case. The questions we would like to answer are:

- How much can load balancing improve system wide Resource Utilization?
- How large should the hysteresis margin and LBC cycle for the load balancing algorithm be in the simulated environment?
- Does load balancing have the potential to release enough resources for rescue handovers and other higher priority traffic or should the usage of guard bands be considered?

Although the results apply only to this specific simulation setup, they will still provide valuable preliminary information on the behavior of the algorithm in Mobile WiMAX.

5.1 System Model and Configuration in the Simulator

In this section we will take an overall look on how the WiMAX system is modeled and configured and what kind of an environment in terms of traffic and topology will be simulated.

[1]The enhanced algorithms presented in the enhancement part will not be evaluated, but could be the target of further research.

The WiMAX system and its environment will be simulated with NS2 [ns2], a discrete event simulator based on two languages, C++ and Object TCL (OTcl). In NS2 C++ is used for packet processing to guarantee fast execution of the simulation and OTcl for control purposes to enable agile configuring. In our simulations we will use shell scripts to run multiple simulation scenarios and use awk, perl and Matlab for post processing the results.

5.1.1 WiMAX system configuration

Here we will go through the configuration, simplifications and working assumptions for the IEEE 802.16e PHY and MAC implementation and the modeled WiMAX Forum network architecture in terms of load balancing.

5.1.1.1 IEEE 802.16e PHY and MAC

The PHY implementation supports a TDD OFDMA frame structure that models most of the elements and dedicated channels presented in Figure 2.1. DL- and UL-MAPs are modeled (with MAP Information Elements (MAP IE)), UCD and DCD messages are transmitted every 2 seconds in the DL subframe and contention and ranging channels are implemented in the UL subframe. A fixed downlink/uplink subframe ratio of 2:1 is used and therefore Resource Utilization will be calculated from the subframe that has more load.

The modeled MAC layer functionality features a Deficit Weighted Round Robin (DWRR) scheduler in the downlink and an Weighted Round Robin (WRR) scheduler in the uplink with support for bandwidth requests [Shr96]. UGS, ertPS and BE services will be used and the scheduling scheme will prioritize higher priority classes over lower classes (e.g. UGS based VoIP over BE based HTTP). Fragmentation and packing of the incoming packets of the service flows is simulated, as well as MSC grouping where service flows using the same MSC will be aggregated to reduce overhead and make slot utilization in the frame more efficient. A simple admission control scheme is modeled, where checks are made to the new arriving service flows whether minimum reserved traffic rate and sufficient delay can be guaranteed. Link Adaptation and power control are not modeled and we will assume that the MSs are static and have fixed Modulation and Coding Schemes.

BS initiated handovers are supported on the MAC level to enable the simulation of load balancing based directed handovers. Simple hard handovers are modeled, without the use of pre-association to the Target BS or context transfers between the Base Stations. This will make the interruption time experienced by the service flows longer than in reality (especially for VoIP connections) but since we are mostly interested in Resource Utilization issues, this won't be limiting issue.

Overall the implemented model represents the most important functionalities of

the PHY and MAC layer and should therefore represent them realistically in our load balancing simulations. A more detailed description of both PHY and MAC configuration can be found from Appendix A in Tables A.1 and A.2.

5.1.1.2 Load balancing

The load balancing framework will be modeled by periodically broadcasting the basic Spare Capacity Report described in part 2.2.2.2 between the neighbors. The SCRs won't be aggregated meaning that the model will correspond to ASN profile C[2].

When load balancing is initiated, HO_req and HO_rsp messages corresponding to the BS initiated directed handovers will be sent between the Serving and Target Base Stations. To simplify our simulations no admission check will be made with the HO_rsp here and since we will simulate only static MSs, no differentiation between rescue and directed handovers will be made.

The logic of the basic load balancing algorithm is implemented as described in 4.1.2. The default value for the Load Balancing Cycle length will be 1 second since it is defined as a default in the WiMAX Forum network architecture and as the default value for the hysteresis margin we will use 10 % as was used in [Vel04]. Later we will run two simulations cases where we will evaluate the behavior of the system when different values for these parameters for the basic load balancing algorithm are used.

We will use a ready configured lists of MSs in overlapping areas (cell re-selection process won't be modeled) so that the BS will be able to handover MSs right away after load balancing is enabled. As discussed in part 2.2.2.2, the BS controlled load balancing handovers will only be conducted for non-BE connections. One MS per frame will be handed over, so no parallel handovers will be made. As was mentioned in the description of the basic algorithm, no prioritization of the order in which the MSs are handed over will be done.

All BSs in the system will have the same capacity and configuration and the MSs can therefore use that same UCD and DCD information when re-entering the TBS acquired in the first enter to the system. The MOB_NBR-ADV message won't be broadcasted at all.

5.1.2 Environment

In the following we will describe the environment, in which the above described WiMAX system will function. We will use a simplified access network topology and MS distribution that will model the congestion of a BS in relations to its neighbors. The channel will be modeled with fixed MCSs and the traffic will be a mixture of

[2]The results should apply quite well also for ASN profile A, since the only difference will be in the way the SCRs are delivered to the BSs.

User Datagram Protocol (UDP) based VoIP traffic with guaranteed throughputs and TCP based elastic FTP and HTTP traffic with BE service.

5.1.2.1 Topology and channel

We will use a simplified system model by modeling our access network as a cluster of three Base Stations with omni-directional antennas. The BSs will reside side by side with the overloaded BS (BS 2) in the middle.

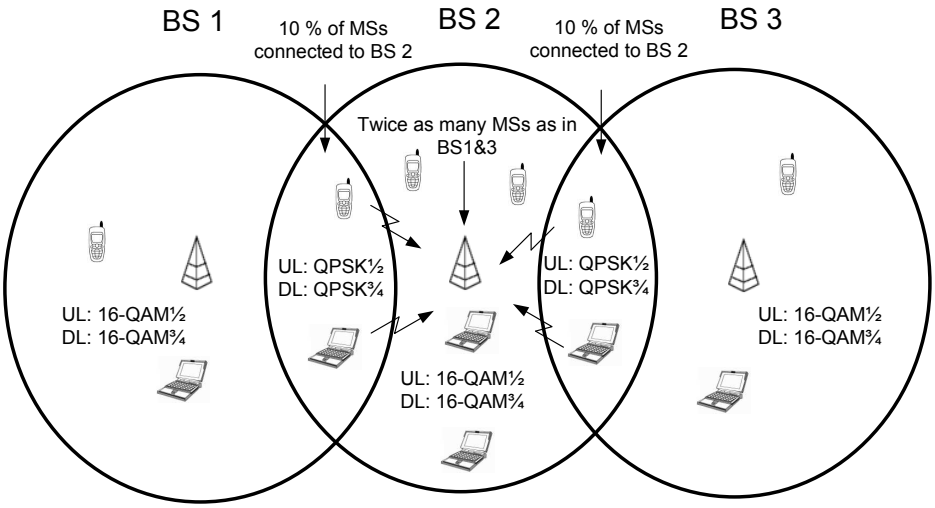

Figure 5.1: Simulation setup.

The MSs will be distributed to the system so that twice as many MSs will be dropped to the middle BS than to the less congested neighboring BSs (BS 1 and BS 3). The total number of MSs dropped to the system will be 400 meaning that 200 MSs will be dropped to BS 2 and 100 respectively to BS 1 and BS 3 (see Table A.7 for the specific MS distribution). Detailed interference, shadow and fast fading models won't be used since they are not an important issue in terms of load balancing for static nodes with rather static channel conditions. We will instead model the channel by choosing appropriate fixed MCSs.

As can be seen from Figure 5.1 the MSs residing in the overlapping areas far away from the BS will use a more robust MCS than MSs residing closer to the BSs thus modeling the effect of path loss in the channel. Furthermore a more robust MCS will be used in the UL than in the DL as is commonly done in radio communication systems.

To simplify our simulations and congest the middle BS more we will distribute the MSs so that only MSs connecting to the middle BS will be dropped to the overlapping area and all the MSs connecting to the lightly loaded BSs will be dropped closer to the BS. The MSs dropped to the overlapping areas (10 % + 10 % of the

total number of MSs connecting to BS 2)[3] will be handed over to the less congested BSs if necessary.

Although this setup doesn't necessarily reflect a very realistic MS distribution, it should still be sufficient for the preliminary evaluation of the basic algorithm. Detailed values of the MCS and the default MS distribution can be found from Appendix A in Tables A.5, A.6 and A.7.

5.1.2.2 Traffic and QoS

Traffic used in the system will be a mixture of four traffic types VoIP, VoIP with Voice Activity Detection (VAD), FTP and HTTP which will be served by the scheduler with UGS, ertPS and BE services. In our simulator we will use existing NS2 traffic generators for the first three traffic types and a separately implemented traffic generator for HTTP based traffic.

The traffic type with the highest priority will be UGS based VoIP, whose generator will send fixed size data packets on a fixed periodic interval. To simulate a bidirectional connection a traffic generator will be configured for both the UL and the DL.

The next highest priority traffic type will be ertPS based VoIP with Voice Activity Detection (VAD). The generator for VoIP with VAD will generate traffic during an activity period in a similar way as regular VoIP, but during a silence period nothing will be transmitted.

The lower priority BE traffic types will be FTP and HTTP traffic. FTP traffic will be simulated with the NS2 inbuilt FTP traffic generator, that sends data constantly according to the TCP congestion window. The implemented HTTP traffic generator will send packets according to the model specified in [3GPP2] where the process of downloading a web page consists of main and embedded object retrievals and the corresponding reading and parsing times. A more detailed description of this can be found from [Cas07]. For FTP and HTTP data transfer is asymmetrical since only acknowledgments will be sent in the UL.

Each MS will carry only one UL&DL service flow pair and the traffic distribution among the MSs will be equal (25 % of the MSs use VoIP, 25 % VoIP with VAD, 25 % FTP and 25 % HTTP). As mentioned earlier load balancing will only be conducted for non-BE connections meaning that with this traffic mix load balancing will only be conducted for VoIP and VoIP with VAD. More information on the traffic parameters can be found from Appendix A in Tables A.3 and A.4.

[3]As discussed in part 2.1.1.2 in Mobile WiMAX a frequency reuse factor 3 will be used in the edges of the cell limiting the size of the overlapping area and thus the number of MSs in the overlapping area. However some overlap will also occur between the BS sectors and hence this assumption should be reasonable.

In [Job04] an interesting study was presented on how and why hot-spots happen. Three typical cases were identified: delay based, capacity based and preferential mobility based hot-spots. In the delay based case the time that moving MSs spend in the cell[4] increases (e.g. traffic jam). In the capacity based case, the capacity of the BS is temporarily reduced (e.g. node updates)[5]. In the preferential mobility based case people are moving towards an event (e.g. a concert) and hence the number of users in the BS increases[6].

In our simulations we will congest the middle BS based on preferential mobility and delay. Preferential mobility will be simulated by dropping more MSs to the middle cell. The delay case will be simulated by not terminating the initiated service flows during the simulation which will also simplify implementation. Flow arrivals will be modeled with a Poisson arrival process with an average service flow inter-arrival-time of 1.2 seconds (0.83 flows per second arrival rate)[7]. We can use Little's formula (4.3) to show that this is a valid arrival rate to congest a BS. In [Die04] a 180 second call holding time was used to model a typical length of a call but since data sessions usually last longer, a flow with this profile could last on average let's say 300 seconds. If this holding time is used in conjunction with the chosen flow arrival rate, Little's formula will result in $0.83 \cdot 300 = 250$ connections on average in the BS which would clearly overload a BS in the system.

5.1.3 Evaluation cases and measurements

Altogether there will be three evaluation cases. In each simulation case we will run several simulations and compare the results obtained from each individual simulation with each other. Individual simulations will be run only once and hence the results cannot be assumed to be general. However the same traffic and arrival process will be generated for all simulations making them comparable in relations to each other.

In the first case we will run two simulations, one where the basic load balancing algorithm will be used and another where no load balancing is conducted. The idea is to get a general idea of how the algorithm works and see how much it can improve system wide Resource Utilization.

In the second and third evaluation case we will study how the algorithm behaves in the configured environment as a function of its two parameters, the size of the hysteresis margin and the length of the Load Balancing Cycle.

[4] Corresponding to the average holding time t_s in Little's equation (4.3).

[5] This could be compared to the reduction of capacity experienced by new arriving MSs when handover guard bands are used.

[6] Arrival rate λ_{res} in equation (4.3) increases.

[7] More extensive simulations with more detailed arrival and departure processes could be conducted in the future.

The use of a very small hysteresis value (0 % or 5 %) should cause a *handover based* ping-pong effect. Since the size of the hysteresis margin increases as the average load increases the maximum evaluated hysteresis value will be dimensioned so that it will not pass the maximum capacity. In terms of the LBC length, in our rather static evaluation environment, it is expected that a rather long value will be sufficient since the packet level non-BE traffic does not fluctuate much and flow level traffic increases and decreases are rather slow.

In all simulation cases we will first load the system in a balanced way by dropping all MSs destined to the lightly loaded BSs (BS 1 and BS 3) and by dropping the same number MSs to the middle BS. After the system is approximately in balance we will start overloading the middle BS by dropping more MSs to it, activate the basic load balancing algorithm and observe how the system behaves. The MSs will be dropped randomly according to the defined spatial distribution (See Figure 5.1 and Table A.7) and arrival process.

Performance of the algorithm will be evaluated with instantaneous[8] and average results. The aim of the algorithm is to balance the load of non-BE traffic and therefore we need to track the non-BE Resource Utilization of the BSs. An easy way to see the load unbalance in a system is to use the system load balance index defined in equation (3.2).

To illustrate how the system behaves we will measure and plot the UL and DL subframe Resource Utilization of both non-BE and BE data and also the Resource Utilization of the subframe headers that carry e.g. the DL- and UL-MAPs in the DL subframe and contain the contention and bandwidth slots in the UL subframe[9].

As already discussed, the system also needs to tolerate some unbalance to avoid unnecessary "ping-pong" handovers. These will occur if the hysteresis margin is set too small. We will evaluate this by recording the total number of handovers conducted during the simulation and what's more important the maximum number of handovers experienced by a single MS.

On the other hand if too much unbalance is tolerated Resource Utilization and reservation will grow to a high value and admission control will start to block incoming calls to protect the QoS of the existing non-BE flows and the service received by the BE flows will start to degrade (longer delays and lower throughput). Hence these effects will also be measured.

[8]The averaging interval for the instantaneous results will be 1 second.

[9]Note that the non-BE Resource Utilization reported in the SCR will be a sum of both non-BE data and subframe headers.

5.2 Simulation results

Now that we have a clear understanding of the simulation environment we can run the evaluation cases and see how the basic load balancing algorithm behaves in the simulated environment. Here we will first present results from each simulation case and then have an overall discussion of the results.

5.2.1 Results from each evaluation case

In the following three parts we will go through results obtained from the three simulation cases: load balancing activated versus inactivated, hysteresis margin evaluation and Load Balancing Cycle length evaluation.

5.2.1.1 With LB vs. without LB

Load balancing with handovers has the potential to improve system wide Resource Utilization by distributing the load to the less congested BSs in the system. Figures 5.2 a-b present overall results from the first evaluation case consisting of two simulation runs, one where the basic load balancing scheme was used and another where no load balancing was conducted.

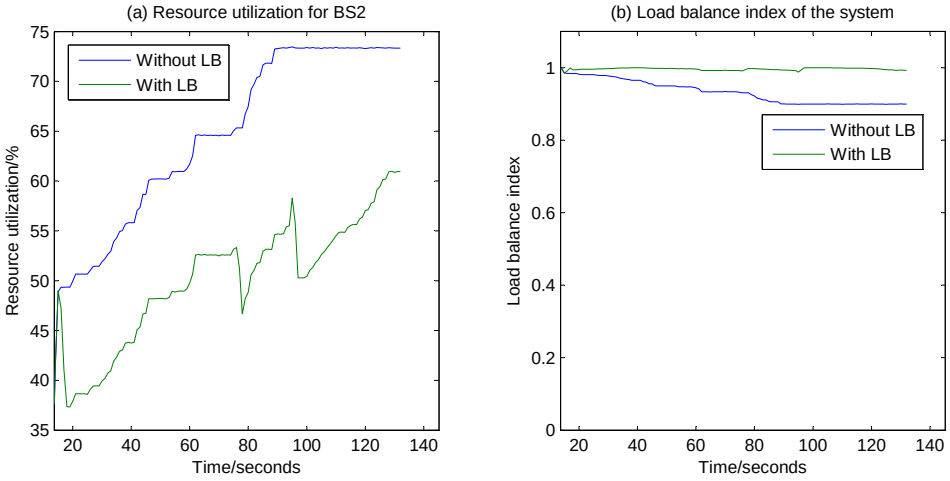

Figure 5.2: Resource Utilization of BS 2 (a) and the load balance index (b) with and without load balancing.

When load balancing was not used no directed handovers were conducted (no Resource Utilization decreases in BS 2) and as a result admission control had to block 19 new non-BE (VoIP based) calls in the congested BS 2 whereas with the basic load balancing algorithm the non-BE load was distributed to the other BSs and no new calls had to be blocked in BS 2. Also we can observe that the load balancing algorithm was able to keep the load balancing index close to the target value 1 throughout the simulation whereas without load balancing the index resulted in a

value of 0.9 even when a large portion of the non-BE load was blocked.

To get a better understanding of what actually happens in the system, the Uplink and Downlink Resource Utilizations for all three Base Stations are depicted for the simulation run without load balancing in Figures 5.3 a-f and for the simulation run with load balancing in Figures 5.4 a-f.

The Figures 5.3 a-f and 5.4 a-f present four Resource Utilization curves: one for subframe headers corresponding to the DL- and UL-MAPs in the DL subframe and the contention and bandwidth slots reserved in the UL subframe, one for non-BE data corresponding to the VoIP based data traffic (including corresponding MAC headers and management messages), one for BE data corresponding to the FTP and HTTP data in the DL and acknowledgments in the UL and one for the total Resource Utilization[10].

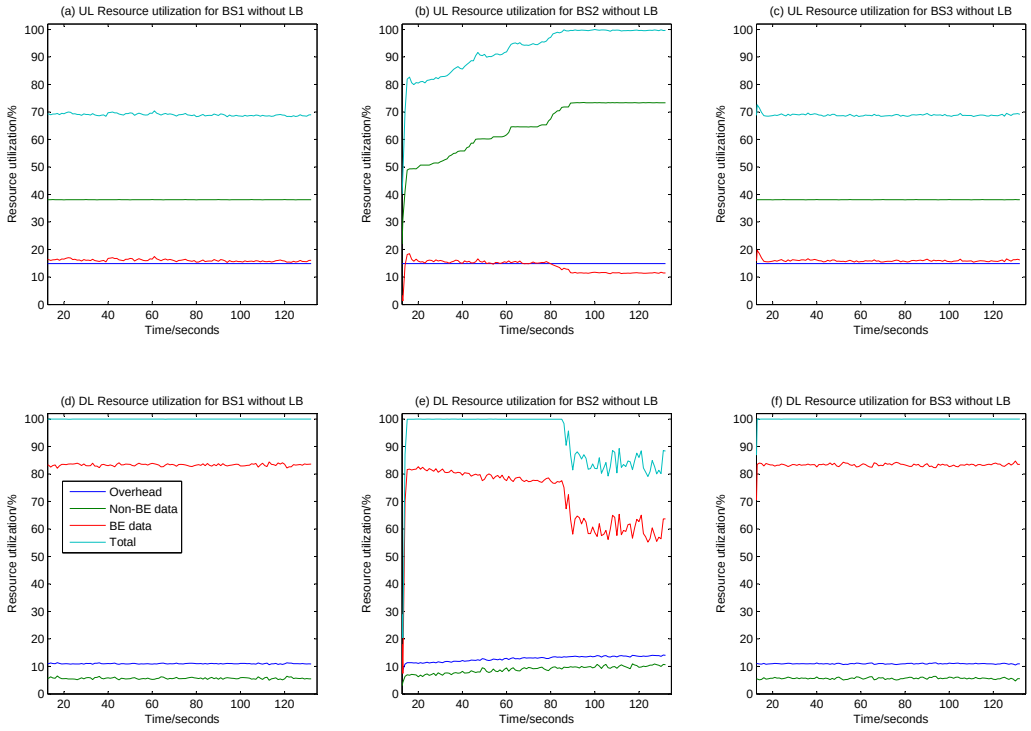

Figure 5.3: UL (a-c) and DL (d-f) Resource Utilizations when no load balancing used.

As can be seen the UL subframe is clearly the bottle neck for non-BE traffic and will determine the final Resource Utilization value as defined in equation (3.3), used to trigger load balancing. We can see that as more MSs are dropped to BS 2, the non-BE data Resource Utilization gradually increases until admission control starts

[10]Note that the Resource Utilization reported in the Spare Capacity Report will be a sum of header and non-BE data Resource Utilization.

to block new VoIP based flows. This happens because a limit in the uplink Resource Reservation has been reached after which the QoS of the existing VoIP calls would degrade. The 19 non-BE VoIP based calls blocked result to a 19 % blocking rate in BS 2 during the simulation, as 100 VoIP based MSs were dropped to BS 2. Despite of the call blocks we can see that at the end of the simulation the system is still quite unbalanced with about a 30 % difference in the uplink Resource Utilization.

In addition what is very interesting here is that even though some bandwidth is left for the acknowledgments of BE traffic in BS 2, the slight decrease in the UL BE throughput results in quite a large drop in the downlink BE throughput.

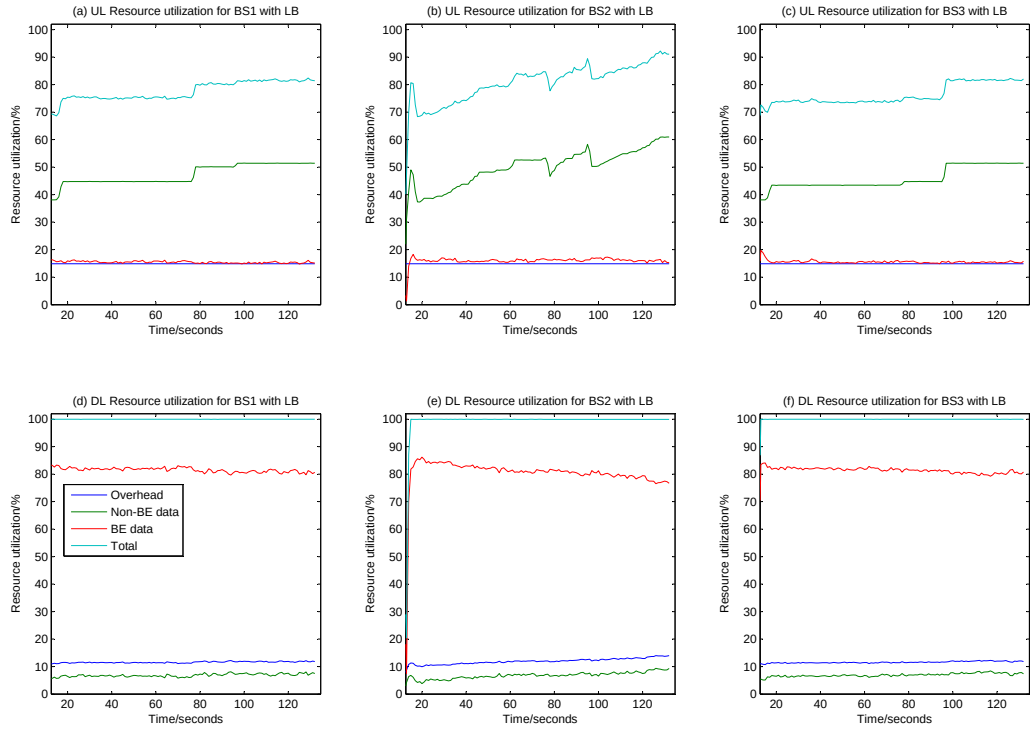

Figure 5.4: UL (a-c) and DL (d-f) Resource Utilizations when load balancing is used.

When the same simulation is run with load balancing we can see that as the non-BE Resource Utilization increases the load is distributed to the less congested Base Stations, BS 1 and BS 3 and hence no calls need to be blocked. A 10 % hysteresis value is used, so each time the Resource Utilization of BS 2 surpasses the average Resource Utilization in the system with 10 percent of the average, the BS initiates directed handovers for the non-BE MSs in the overlapping areas. In this simulation run load balancing is initiated three times at around 16, 76 and 95 seconds. The directed handovers can be seen as three steep drops in the Resource Utilization of BS 2 and corresponding increases in BS 1 and BS 3 (clearly in the UL and to some degree also in the DL).

We can see that at the end of the simulation run the system is quite well balanced. The slight unbalance is due to the hysteresis margin. What can also be observed is that since the BE connections in BS 2 have enough bandwidth to send the acknowledgments the BE DL Resource Utilization doesn't decrease as in the simulation run without load balancing, but is here reduced only slightly as non-BE data Resource Utilization increases. We can recognize this difference also in the total system DL FTP throughput presented in Figure 5.5.

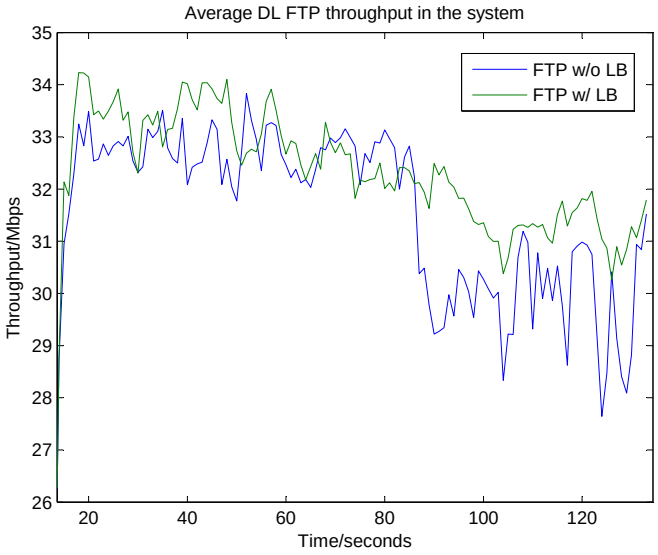

Figure 5.5: System wide downlink FTP throughput with and without load balancing.

From the Figure 5.5 we can see that after the UL gets too congested in the simulation run without load balancing, a clear decrease in FTP downlink throughput can be seen in the whole system when compared to the simulation run where load balancing is used.

This simulation case has shown that load balancing can in fact have a rather large impact on the system wide Resource Utilization in case of congestion and is therefore a valid method. In the next part we will evaluate the behavior of the system as a function of the hysteresis margin.

5.2.1.2 Evaluation of the hysteresis margin

The hysteresis margin is set to tolerate some unbalance in the system and hence avoid the handover or fluctuation based ping-pong effect discussed in 4.1.3.1[11]. While it is good to avoid temporary overloading and unnecessary handovers there is a limit to how large the hysteresis margin can be and how much unbalance can be tolerated. In this evaluation case we will observe how the system behaves as the size of the load balancing margin is increased.

In Figures 5.6 a-c results from the simulation runs are summarized in the form of the number of blocked calls, average load balance index, and both the total number of directed handovers conducted in the system and the maximum number of directed handovers experienced by a single MS.

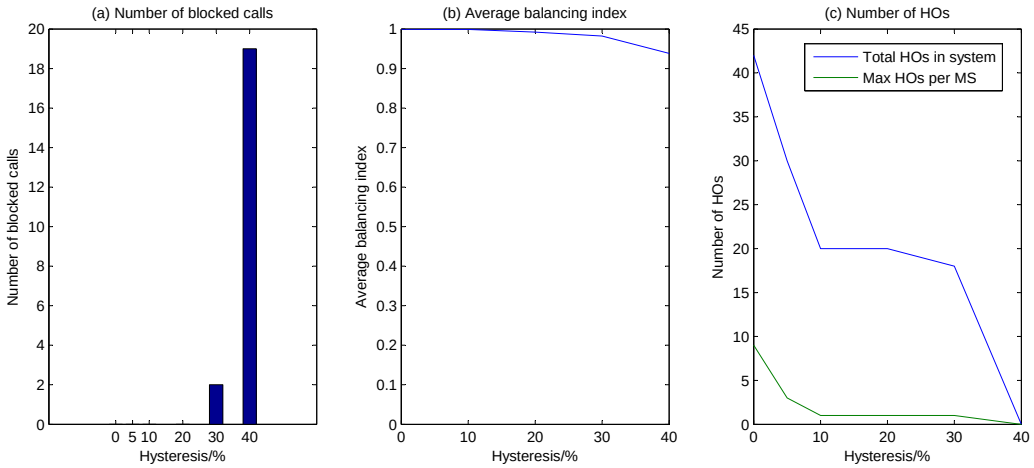

Figure 5.6: Number of blocked calls (a), average load balance index (b) and number of directed handovers (c) with different hysteresis margins.

Overall we can see that as the size of the hysteresis margin increases, the number of blocked calls increases (or stays the same) and the number of directed handovers conducted decreases (or stays the same). At the other end a 0 or 5 % hysteresis margin seems to be too small since it results in a handover based ping-pong effect and at the other end a 30 % or 40 % hysteresis margin too large since it results in new call blocking. With a 10 % and 20 % hysteresis margin no new call blocking or unnecessary handovers occurred. The average load balancing index in the middle shows that unbalance grows as the size of the hysteresis margin is increased. Figures 5.7 a-b illustrate in more detail the instantaneous behavior of the system with each evaluated hysteresis margin value.

[11]Since the non-BE traffic served in the simulations is rather static only the handover based ping-pong effect will appear.

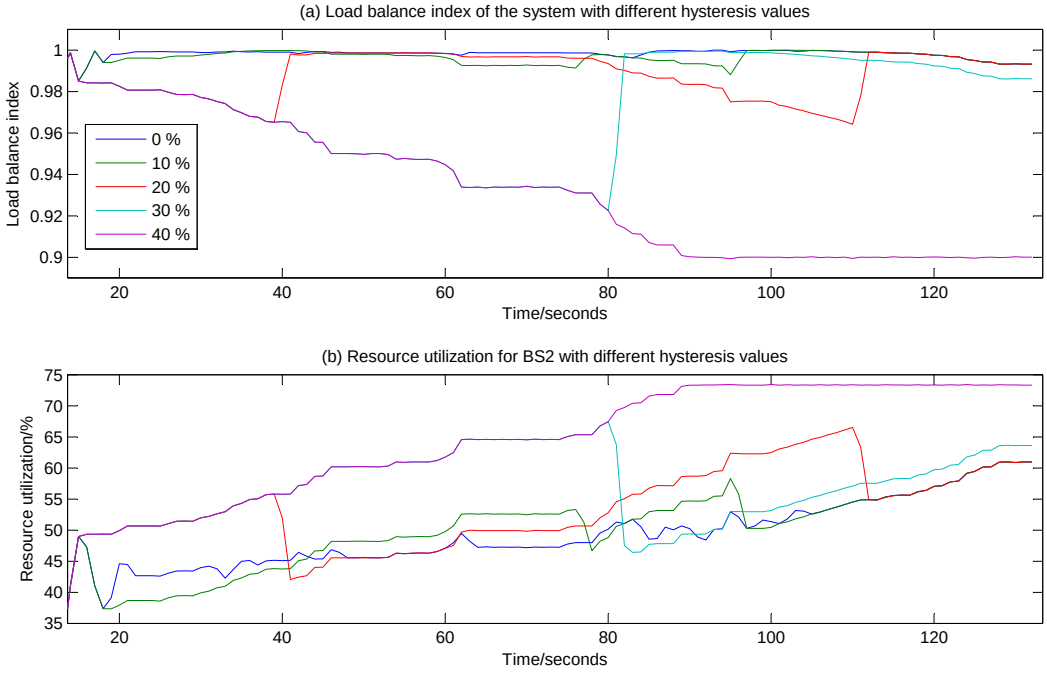

Figure 5.7: Instantaneous load balancing index (a) and Resource Utilization of BS 2 (b) with different hysteresis margins.

Both the load balance index and the Resource Utilization of BS 2 describe well how the load is balanced with the different hysteresis values. As can be seen the larger the hysteresis the longer the load balancing algorithm will wait before reacting to the traffic increase and distributing the load so that the system is in balance again (balance index 1). Load balancing is shown as steep increases in the balance index and steep decreases in the Resource Utilization[12].

As can be seen with a 0 % hysteresis margin the load balancing index stays very close to 1 at the expense of ping-pong handovers coming back from BS 1 and BS 3. They can be seen as steep increases in the Resource Utilization of BS 2 that deviate from the arrival pattern of the other curves (seen at about 18, 33, 85, 91 and 100 seconds). A 10 % margin seems to be enough to avoid the handover based ping-pong effect since then all 20 non-BE MSs residing in the overlapping areas are handed over only once with load balancing being triggered three times in BS 2.

When a 20 % hysteresis was used load balancing was triggered twice and the end result was the same as with a 0 and 10 % hysteresis. When a 30 % hysteresis was used load balancing was triggered only once. However with a 30 % hysteresis two

[12]At the end of the simulation the system becomes unbalanced because all MSs residing in the ovelapping areas have been handed over to the less congested BSs and hence no more directed handovers can be made.

calls were blocked just before load balancing was triggered meaning that this margin is close ideal in this particular case but is still a bit too large. The curve for the hysteresis margin 30 % in Figures 5.7 a-b deviates from the others because two VoIP based MSs arrived to the system after load balancing was initiated and no handovers were conducted for them.

A 40 % hysteresis margin proved to be too large since admission control started to block calls before the triggering threshold in Resource Utilization was reached and as a result the triggering threshold was never reacher and no directed handovers were conducted.

In general to avoid such call blocking, Resource Reservation based load balancing triggering (see part 4.2.1.2) should be considered. However with this particular traffic profile where non-BE Resource Utilization doesn't fluctuate much and is very close to Resource Reservation an upper limit to the Resource Utilization based triggering threshold could be sufficient. Using such an upper limit would be necessary also due to the fact that in the basic load balancing algorithm (based on [Vel04]) the hysteresis margin is set manually and hence the triggering threshold can, as the average load of the system increases, grow to a value larger than the total capacity. Since VoIP based flows were blocked when Resource Utilization was at about 88 % (74 % non-BE data and 14 % subframe headers), the triggering threshold upper limit in this case could be set for example to about 84 %.

In addition the use of a method that dynamically tunes the triggering threshold according to the state of the system (see 4.1.3.1) should be considered since the manually set threshold in the basic load balancing algorithm produces a threshold that is only dynamic in the sense that the size of its hysteresis margin increases as the average load increases and therefore might not meet the needs of the system with more mixed and fluctuating traffic profiles.

Another interesting aspect that came forth from the hysteresis based simulation case was the issue of estimating how many MSs should be handed over by the SBS and accepted by the TBS. As specified in the basic load balancing algorithm (see Figure 4.2) directed handovers were conducted until enough resources were released so that the system wide average was reached. In general, especially if the hysteresis margin is quite small, this should be done with care as handing over too many MSs might cause a too large of an increase in Resource Utilization in the Target BS resulting in the handover based ping-pong effect.

Although with this particular traffic profile it was quite simple to estimate how much of the resources will be released in SBS and increased in TBS when an MS was handed over, when the traffic and channel are more varying, it can be very

challenging[13]. This problem can be mitigated to some degree by using most up to date Resource Utilization measurements when the decision to accept the directed handover is made. However since these estimations are equally difficult to make in the TBS it might be better to be a little conservative and stop handing over the connections before the average level is reached to further avoid the ping-pong effect and make load balancing more gradual[14].

So what hysteresis margin should be chosen for this traffic profile? Even though a 10 % hysteresis already eliminates the handover based ping-pong effect, it might be better to set it to a more conservative value (say little over 20 %) due to the fluctuations that will come from the varying channel and MCSs changes. Since VoIP calls only need a certain guaranteed throughput (with delay requirements) and since they won't benefit from extra bandwidth, one could argue that with this particular traffic profile we should set the triggering threshold as high as possible and admit as many VoIP calls as we can as long as an upper limit for the Resource Utilization based triggering threshold (based on scheduling and admission control) would be set. However this could come with the cost of a decrease in the BE performance. In Figure 5.8 we can see how the average delay in the system changes for the different traffic classes in our simulations as a function of the hysteresis margin.

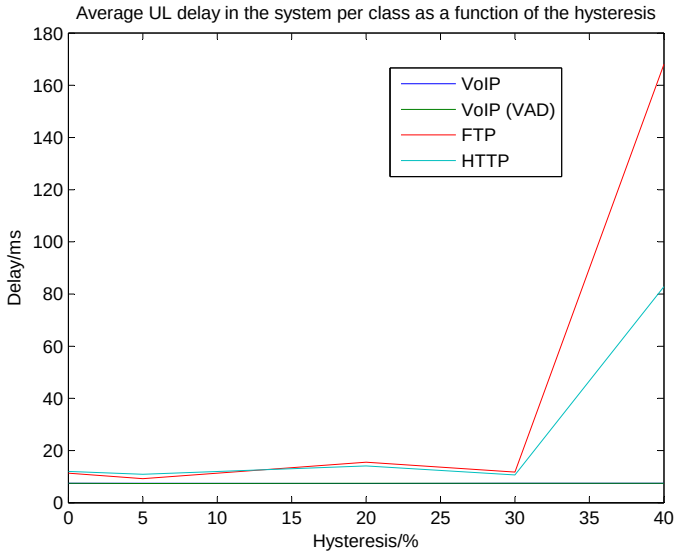

Figure 5.8: System wide delay for the traffic classes with different hysteresis margins.

Both VoIP and VoIP with VAD don't experience any degradation in terms of their

[13]In our simulations with the used UL MCS for the MSs residing in the overlapping area, one VoIP flow (MAC headers and management messages included) consumed approximately 1.3 % of the resources.

[14]This issue of estimating how much of the resources the flow will consume is closely related to admission control and could be studied in conjunction to it.

UL delay since admission control protects them. As we can see the UL delay for the BE traffic types FTP and HTTP increases dramatically with the hysteresis margin 40 %. Although the principle is to prioritize higher priority traffic over lower (i.e. VoIP over BE), this should be done with reason.

Commonly when provisioning bandwidth, a small part of if it will be reserved only for BE traffic to guarantee at least some throughput for it (e.g. [Zha04] suggests 20 % bandwidth reservation for BE traffic). In this case where the decrease of throughput for the BE traffic acknowledgments causes a considerable decrease in the DL BE data throughput using such a guard would be beneficial.

On the other hand, if load balancing with handovers would be supported in the terminals the delay increases experienced by BE FTP and HTTP connections could result in MS initiated load balancing based handovers for the BE MSs (and hence the BE connections would conduct a handover first to the less congested BSs). Furthermore if the additional fields mentioned in 4.1.3.2 would be implemented, also the BS could initiate directed handovers for the BE MSs. This would be better because the BS would have more information and would also know what would be the best TBS for the MS to handover to, in terms of available bandwidth for the BE MSs and the number of other BE MSs contending for it in the candidate TBSs.

So in conclusion with this particular traffic profile, a 20 % margin seems to be good since it is large enough so that handover ping-pong effect won't occur, but on the other hand low enough so that it will not cause call blocking or disturb BE traffic to a high degree. The chosen hysteresis value could be complemented with an upper limit for the triggering threshold for Resource Utilization being set to about 84 %. We can also conclude from the simulations that the delay experienced by the lower priority MSs (here BE) can be considered as a good indicator that the Resource Utilization based threshold should be lowered as was discussed in part 4.1.3.1.

5.2.1.3 Evaluation of the length of the LBC

The length of the LBC defines how often the Spare Capacity Report is sent to the other BSs and hence how fast load balancing can react to traffic fluctuations. The length of the LBC in the basic load balancing algorithm also defines the length of the averaging interval used. For simplicity the reporting procedure was implemented so that the averaging interval and the SCR reporting were synchronized meaning that the measurements made during one LBC cycle were reported to the other BSs during the next LBC cycle and then used for decision making in the beginning of the third LBC[15].

Even with this limitation the traffic load could be balanced even with a load balanc-

[15] As was discussed in part 4.1.2 it would be better to have the averaging interval and reporting procedure in different synchronization and wait as long as possible to send the SCR, so that the most up to date information is reported.

ing length of 10 and 20 seconds before calls were blocked. Figures 5.9 a-b present the results from this simulation case[16].

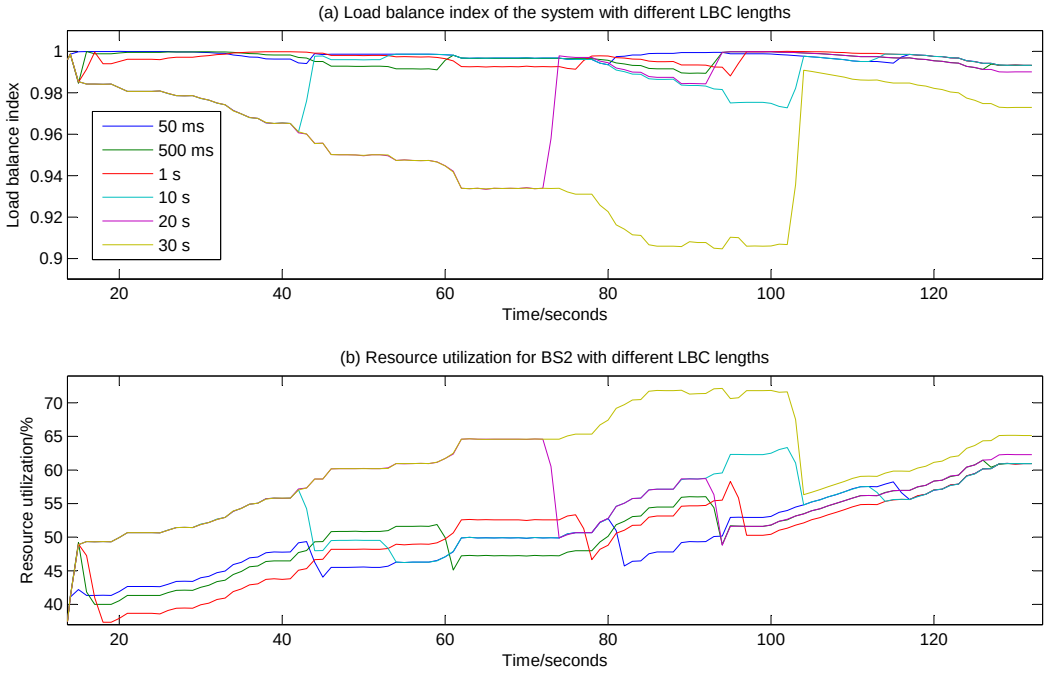

Figure 5.9: Instantaneous load balancing index (a) and Resource Utilization of BS 2 (b) with different LBC lengths.

The simulated LBC length was quite wide ranging from 50 milliseconds (e.g. a 100 ms LBC length was used in [Vel04]) to 30 seconds. As can be seen the longer the LBC duration the longer the algorithm waits until it reacts to the traffic increase.

The balance index remained closer to 1 for the smaller LBC lengths throughout the simulation, but otherwise notable differences was not seen in the performance of the algorithm with the different LBC lengths until the length was increased to 30 seconds and call blocking occured. No ping-pong effect was seen with any of the LBC lengths even with the smallest simulated LBC length, 50 milliseconds.

Some but not a clear correlation between smaller LBC lengths and the number of times load balancing was triggered was seen. Load balancing was triggered four times for the 50 millisecond run (at 14, 42, 80 and 115 seconds) and the 500 millisecond run (at 14, 57, 92 and 126 seconds). As the LBC length was increased a 1 second LBC length triggered load balancing three times and a 10 second LBC length four times. Surprisingly even with a 20 second LBC length where load balancing

[16]The default hysteresis margin 10 % was used.

was triggered only twice no new call blocking occurred[17]. Finally with a 30 second LBC length 7 calls were blocked before load balancing was triggered and the rest of the arriving new calls were able to be accepted.

Even though the load balancing algorithm was able to balance the load for this rather static non-BE traffic profile even with a 20 second LBC length (with the implementation simplification making the simulated algorithm even slower to react) a lower value should still be used since some flow level fluctuation might happen due to MCS changes. In any case as a conclusion from the simulation case we can say that with this traffic profile a rather long LBC period is sufficient. What can also be concluded is that the more static the traffic and channel are, the less important the size of the averaging interval is since the performance of load balancing stays the same whether we have a very short averaging period or a rather large SCR reporting period. Therefore the default value of 1 second used in the WiMAX Forum network architecture seems a pretty reasonable choice also for this specific traffic profile.

If we take this issue a bit further, it is quite interesting that, as traffic fluctuation increases, the choice of a good LBC length becomes very difficult because we should on the other hand make the averaging length larger to be able to report better averaged results but on the other hand shorten the SCR reporting interval so that load balancing will get most up to date information and is able to react to the varying conditions. In this sense it might be better not to change the LBC length too much and but increase the hysteresis margin as the traffic becomes more fluctuating[18].

Another interesting aspect that came up during the simulation case was that it is important to use measurements from the same LBC when calculating the average Resource Utilization and also to make sure that they reflect the actual loading situation in the system. In [Vel04] the interruption time for the handovers (re-associations in WLAN) was quite long and hence no average load was reported during the interruption time but old values were used.

Since in our simulations the interruption time during the handover was in almost all cases less than 100 ms and since the proportion of one VoIP based MSs of the whole load was not that significant, the calculated average value did not change much. However if the interruption time would become much larger and the system load would decrease temporarily in a considerable way, the SCR report should not be sent and old values should be used to calculate the average.

[17]With a 20 second LBC length the last load balancing triggering happened before the last MSs were dropped to the overlapping area and hence their Resource Utilization ended up a little higher.

[18]This could be complemented by a so called time hysteresis where a certain *hysteresis time* could be set for how long the Resource Utilization should remain over the triggering threshold once it has passed it, before load balancing is triggered [Sol06]. The hysteresis time could also be longer for the delay sensitive connections if multiple thresholds are used.

What is also very interesting in relations to receiving correct measurements from the neighboring BSs, is how the BSs will be synchronized in relations to each other. For the algorithm to work properly, accurate measurements are needed meaning that the BSs should be synchronized at least to some degree. The neighboring BSs should measure their load around the same period of time and also report their loading situation around the same time so that correct decisions can be made. Changing the LBC length without losing synchronization might be a difficult procedure especially if there is no centralized element controlling the whole (i.e. ASN profile C). This is another aspect that contributes to the reasoning to use the same default value for the LBC length and react to the traffic by changing the hysteresis margin. How this BSs synchronization should be done could be the target of future research.

What was also seen in the simulations was how fast load balancing with handovers was able to release resources (very steep decreases in Resource Utilization when load balancing triggered)[19]. This increases the ability for load balancing to react to traffic changes faster and compensates for the slow reaction of a long LBC length.

5.2.2 Conclusions from the results

Although the evaluations were conducted in a static environment with rather static traffic, we were still able to get valuable information of the basic characteristics of the basic load balancing algorithm and load balancing in general.

All in all the simulations showed that load balancing can be a very efficient way to enhance system wide Resource Utilization in Mobile WiMAX. In our simulations of a system with one overloaded Base Station, the basic load balancing algorithm was able to avoid call blocking altogether (versus 19 calls being blocked when no load balancing was used), improve BE throughput and distribute the load across the system quite nicely.

A clear need to optimize the size of the hysteresis margin was seen as a too small hysteresis caused a *handover based* ping-pong effect and a too large hysteresis caused call blocking and a drop in BE throughput. This could be addressed in the future by further developing the scheme discussed in 4.1.3.1 that automatically tunes the triggering threshold in relations to changing traffic. Furthermore when Radio Resource Fluctuation in the system, F_{sys}, increases *fluctuation based* ping-pong effect is likely to occur which could also be addressed with automatic threshold setting and by further developing the multiple threshold scheme discussed in 4.1.3.3.

One problem with the manually set hysteresis value is that the hysteresis is set as a certain percentage of the average loading level, meaning that the hysteresis margin size increases as average loading level increases and hence the triggering threshold can become larger than system capacity. This also means that with very

[19]Note, however that no cell reselection procedure was simulated and ready lists of overlapping cells was used.

low average Resource Utilization value L the hysteresis margin will be very small.

Therefore, to make the basic load balancing algorithm feasible for deployment an upper and lower limit for the Resource Utilization based triggering threshold could be set. As discussed earlier the upper limit $T_{U,bas,max}$ should be based on scheduling and admission control so that load balancing will be triggered before call blocking occurs and service degrades in the congested BS. The lower limit $T_{U,bas,min}$ could be set based on when load balancing starts to be beneficial since there is no use to balance the load and cause unnecessary handovers if all flows are getting appropriate service.

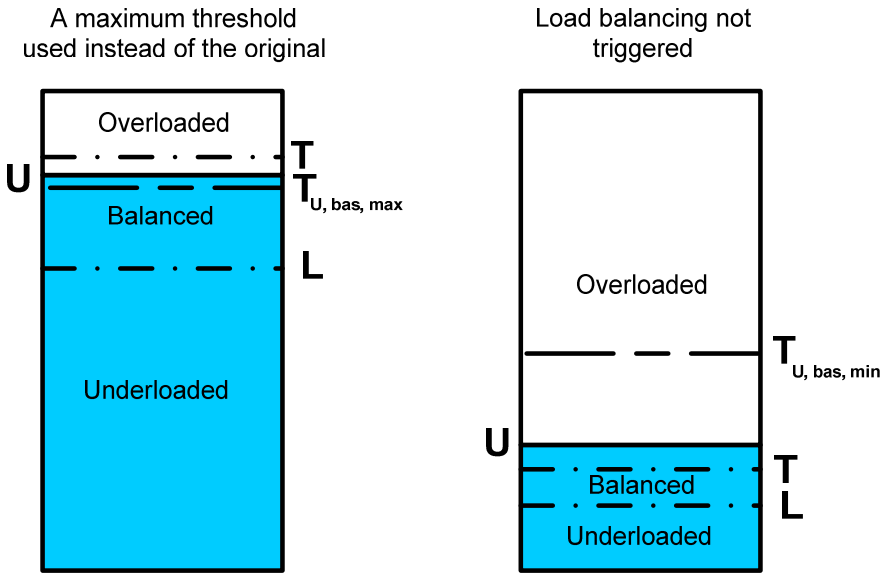

Figure 5.10: Upper and lower limits for the triggering threshold in the basic load balancing algorithm.

As can be seen from Figure 5.10 when the computed threshold is between these limits it will be used, but if the computed threshold falls outside these limits the corresponding limiting value will be used instead. To avoid a handover based ping-pong effect in a situation where the average load L is very close to the upper limit $T_{U,bas,max}$ a maximum average Resource Utilization value L_{max} should be set after which no load balancing would be conducted.

Based on our simulations we concluded that with the used traffic profile, scheduling and admission control, the upper limit could be set to $T_{U,bas,max} = 84\%$. To guarantee a sufficient hysteresis when average load L is high, the maximum average Resource Utilization value could be set to $L_{max} = 76\%$ corresponding still to a little over 10 % hysteresis value in relations to $T_{U,bas,max}$ which should be enough to avoid the handover based ping-pong effect. The lower limit could be set as high as $T_{U,bas,min} = 60\%$ or even more since no degradation of either non-BE traffic or

BE traffic could be seen until the admission control limit was reached. When the 20 % hysteresis value that was concluded to be good in the simulations, is complemented with these boundary values the basic algorithm as such could be deployable.

The simulations showed that the basic load balancing algorithm was able to balance the load for this rather static non-BE traffic profile even with a 20 second LBC length. As Radio Resource Utilization becomes more fluctuating, choosing an appropriate averaging length and SCR reporting period so that on the other hand reliable results are communicated but on the other hand so that the system is also able to react quickly enough to the traffic changes, will become more challenging. As a starting point the default value of 1 second used in the WiMAX Forum network architecture seems a pretty reasonable choice from where more careful tuning could be done.

The results from the simulations can also be used as a preliminary indicator to whether load balancing alone has the potential to release enough resources for incoming rescue handovers or whether a dedicated guard band should be reserved for them to avoid call drops. In the simulations we did see a substantial decrease in Resource Utilization when load balancing was triggered and therefore it could in some cases release enough resources so that all incoming handovers to the BS can be accepted.

However preferential mobility based congestion could be even heavier than in the simulated case (where the BS 2 was twice as congested as the lightly loaded BSs) and also the total load offered to the system could be higher. In addition the size of the overlapping area and hence the number of MSs residing in the overlapping area will have an impact on how much resources load balancing can release. All these factors can decrease the impact of load balancing and hence we cannot assume that load balancing alone can guarantee a sufficient amount of resources for the incoming rescue handovers[20].

If dynamic guard bands would be used this would have an impact on load balancing triggering. In such a case a quite straightforward extension to the basic load balancing algorithm would be to set $T_{U,bas,max}$ (and L_{max}) in relations to the handover guard band (when admission control starts to block new calls) in a similar ways as discussed in 4.2.1.2.

[20]The usage of relay stations with IEEE 802.16j might however make this feasible.

Chapter 6

Summary, Conclusions and Future Work

The main goal of this thesis was to examine how load balancing with Base Station initiated directed handovers could be conducted in Mobile WiMAX and what kind of potential it has to enhance Resource Utilization and QoS system wide. An additional goal of the thesis was also to conduct preliminary research on how system wide QoS could be guaranteed for rescue handovers (and higher priority traffic types) in Mobile WiMAX, how this would affect load balancing and how these two approaches could be combined.

As a summary we can conclude that load balancing with directed handovers can be a very efficient way to enhance system wide Resource Utilization and also enhance the possibility to fulfill QoS guarantees in Mobile WiMAX. However since load balancing cannot itself ensure that enough resources are released for incoming rescue handovers in all cases, the use of handover guard bands should be considered.

In the beginning part of the thesis a background study on the key system aspects of the IEEE 802.16e radio interface technology and WiMAX Forum Access Network Architecture in terms of load balancing and handovers was conducted to exhibit the good framework that Mobile WiMAX offers to conduct load balancing between neighboring Base Stations. After that a literary review on load balancing, and system wide handover and traffic prioritization was conducted to get a good understanding of these concepts.

Based on the gained knowledge a basic Resource Utilization based load balancing algorithm tailored for Mobile WiMAX was designed and three enhancement proposals were made. The first defined a framework to automatically tune the load balancing triggering threshold and the second a framework to enable BS controlled load balancing for Best Effort MSs. In the third enhancement a preliminary scheme to trigger load balancing in a fluctuating environment with multiple thresholds was proposed. Its idea is to minimize unnecessary ping-pong handovers for delay sen-

sitive connections and also enable the delay tolerant connections to have access for more bandwidth in a fluctuating environment.

Later a preliminary framework on how to conduct rescue handover prioritization in Mobile WiMAX with a dynamic guard band was discussed. This led to the proposal of a Resource Reservation based triggering scheme where load balancing can be triggered in relations to a reserved guard band. This was further enhanced by a multiple guard band triggering approach where bandwidth is reserved for higher priority traffic. It was concluded that the Resource Reservation based triggering approach complements Resource Utilization based load balancing well and that together they should form a very efficient combination for load balancing.

Finally preliminary evaluation of the basic algorithm in a static environment was conducted. Although the simulations were not extensive, valuable information was obtained of the basic parameters of the algorithm and of the overall performance of the algorithm. The algorithm performed well in the simulated environment and was further complemented with bounding triggering threshold values to make it deployable.

A clear need however was seen for an automatic tuning scheme, such as the one introduced earlier, especially when traffic becomes more fluctuating. Also as traffic fluctuation increases and mobility comes along the use of both the multiple threshold triggering scheme and the Resource Reservation based triggering scheme should be considered. All of these schemes could be further developed and elaborated in the future and evaluated with more extensive simulations.

BS controlled BE load balancing came up as a potential extension but since this would require additions to the WiMAX Forum network architecture specification the possibility to add these changes should be investigated before further development. Another important addition that could be done to the existing specification that came forth was the differentiation between directed load balancing and rescue handovers. If load balancing and handover prioritization would be conducted at the same time their handovers should receive different treatment in the Target BS. What was also proposed was that load balancing directed handovers would be conducted only for MSs that are likely to reside in the overlapping area throughout their session and won't conduct rescue handovers since this will reduce the number of unnecessary handovers and unnecessary scanning.

All in all this thesis should have formed a very good basis for the further development and evaluation of handover based load balancing in Mobile WiMAX. In the future, more elaborate evaluations of the efficiency of the load balancing schemes could be conducted. These could feature the rtPS and nrtPS scheduling services and corresponding more fluctuating traffic and a more realistic arrival and departure process. In addition the impact of mobility, rescue handovers and the corresponding

rescue handover prioritization scheme could be evaluated. Also the simulation of the discussed handover mechanisms that speed up handover execution, such as pre-association to the Target BS, Optimized Hard Handover (with MS context transfer), FBSS and MDHO and the effect of cell-reselection, could give further valuable information on the actual effect that load balancing with handovers has on the system.

Other interesting fields that could be studied more specifically and in conjunction with load balancing in the future are location and velocity estimation (i.e. identifying static/mobile MSs), the effect of transmission power and interference, BS synchronization within the ASN, and admission control and resource consumption. Future enhancements could also feature load balancing from micro to macro cells or even to other parallel systems such as UMTS. Finally, the introduction of relay stations (IEEE 802.16j) to Mobile WiMAX is expected to improve the efficiency of handover based load balancing substantially. It is a very attractive target of research since it might make load balancing so powerful that it could by itself even eliminate rescue handover drops and therefore ensure the fulfillment of QoS system wide.

Appendix A

Configuration

Here we will present the WiMAX system and environment configuration for the NS2 simulations. The different aspects of the configuration can be mapped to the earlier mentioned triangle model and are shown in Figure A.1.

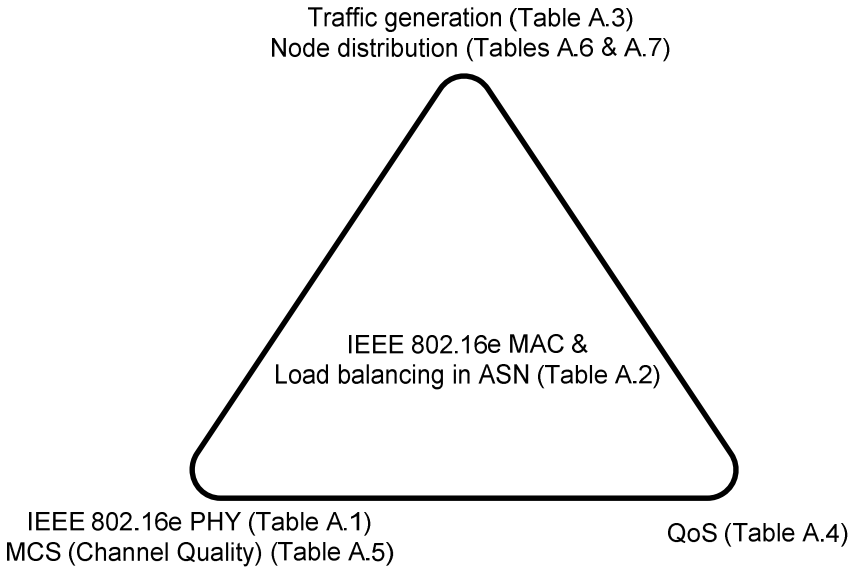

Figure A.1: Configuration mapped to the triangle model.

Table A.1: PHY configuration in the simulator.

Primitive and derived parameters:	
PHY mode	OFDMA
System Channel Bandwidth	10 MHz (frame length 5 ms)
FFT size	1024
Sampling frequency	11.2 MHz
subcarrier frequency spacing:	10.94 KHz
Guard Time	11.4 microseconds
OFDMA symbol time	102.857 microseconds
Number of OFDMA symbols in a frame	48
Resource allocation:	
Subcarrier scheme	PUSC
Uplink Sub-Channels	35
Downlink Sub-Channels	30
DL/UL subframe ratio	Fixed (2:1)
Total PUSC slots per frame	655
DL slots	480
UL slots	175
DL/UL subframe overhead:	
DL/UL-MAP	Modeling based on MAP IEs
UCD/DCD	Sent every 2 seconds
Contention	10 opportunities
Ranging	4 opportunities
Other:	
Coding	Convolutional Turbo Code (CTC)
Antenna type	Omni-directional
Antenna scheme	Single Input Single Output (SISO)

Table A.2: MAC and Load Balancing configuration of the system.

Scheduling:	
Scheduling scheme	(Deficit) Weighted Round Robin ((D)WRR)
Downlink	DWRR (with base quantum 200 for BE)
Uplink	WRR (with base quantum 50 for BE)
MSC grouping	Used
Fragmentation and packing	Used
Admission control	VoIP flows blocked if their QoS cannot be guaranteed
Automatic Repeat reQuest (ARQ)	Not used
Handovers:	
Cell reselection	Not modeled A ready list of MSs residing overlapping areas No pre-association to TBS
HO decision initiation	BS initiated directed handovers
Messages used	MOB_BSHO-REQ MOB_MSHO-IND
HO execution	Complete re-registration done Contention based ranging No context transfer
Default values for load balancing:	
Hysteresis margin	10%
LBC length	1000 ms

Table A.3: Traffic generation.

Traffic generators	**Distribution & Parameters**
VoIP	
Packet size	31 bytes
Packet inter-arrival-time	20 ms
Resulting throughput	12.4 kbps
VoIP with VAD	
Packet size	31 bytes
Packet inter-arrival-time	20 ms
Talk spurt length	Exponentially distributed Mean = 1.026 seconds
Silence length	Exponentially distributed Mean = 1.171 seconds
FTP	
File size	Infinite → data sent constantly according to the TCP congestion window
HTTP	
Main object size	Truncated Lognormal distributed Mean = 10 710 bytes Std. dev. = 25 032 bytes Minimum = 100 bytes Maximum = 2 000 000 bytes
Embedded object size	Truncated Lognormal distributed Mean = 7 758 bytes Std. dev. = 126 168 bytes Minimum = 50 bytes Maximum = 2 000 000 bytes
Number of embedded objects	Pareto distributed Mean = 5.64 bytes Maximum = 53 bytes
Reading time	Exponentially distributed Mean = 30 seconds
Parsing time	Exponential distributed Mean = 0.13 seconds
Protocol stack	
VoIP	UDP/IP
FTP and HTTP:	TCP/IP
TCP segment size:	1000 bytes

Table A.4: Traffic profile and QoS configuration in the simulator.

Traffic applications	
VoIP without VAD	25%
VoIP with VAD	25%
FTP	25%
HTTP	25%
QoS	
VoIP	
UGS MSTR (guaranteed throughput)	12.4 kbps
VoIP with VAD	
ertPS MSTR (guaranteed throughput)	12.4 kbps
FTP	
BE MSTR	256 kbps
HTTP	
BE MSTR	256 kbps

Table A.5: Modulation and Coding Schemes (channel).

Connection	**MCS**	**Capacity**
Overlapping areas (far from BS)		
UL	QPSK1/2	6 bytes/slot
DL	QPSK3/4	9 bytes/slot
Non-overlapping areas (closer to BS)		
UL	16-QAM1/2	12 bytes/slot
DL	16-QAM3/4	18 bytes/slot

Table A.6: Topology (MS distribution).

Topology (MS distribution)	
Total number of MSs in the system	400
Overloading percentage of BS 2	200 %
Proportion of MSs connecting to BS 2 and dropped to the overlapping area	10 % + 10 %

Table A.7: Number of MSs according to the distribution (the MSs that can be handed over depicted in bold).

	Proportion from all MSs	VoIP 25 %	VAD 25 %	FTP 25 %	HTTP 25 %	
BS 1	25 %	25	25	25	25	100
BS 3	25 %	25	25	25	25	100
BS2&BS1 (overlap)	5 %	***5***	***5***	5	5	20
BS2&BS3 (overlap)	5 %	***5***	***5***	5	5	20
BS2 (middle)	40 %	40	40	40	40	160
		100	100	100	100	400

Bibliography

[IE^{3}04] Air interface for fixed broadband wireless access systems. IEEE Standard 802.16, June 2004.

[IE^{3}05] Air interface for fixed broadband wireless access systems - amendment for physical and medium access control layers for combined fixed and mobile operation in licensed bands. IEEE Standard 802.16e, Dec. 2005.

[WiMAX] Mobile WiMAX - Part I: A Technical Overview and Performance Evaluation, August 2006.

[Gud91] M. Gudmundson, "Analysis of Handover Algorithms," Proc. Vehicular Tech. Conf., 1991, pp. 537-42, May 1991.

[Pol96] G.P. Pollini, "Trends in handover design," IEEE Communications Magazine, Volume 34, Issue 3, pp. 82 - 90, March 1996.

[Mou92] M. Mouly and M-B. Pautet, The GSM System for Mobile Communications, 1992.

[Mar99] P. Marichamy, S. Chakrabarti, and S.L. Maskara, "Overview of handoff schemes in cellular mobile networks and their comparative performance evaluation," IEEE VTS 50th Vehicular Technology Conference, 1999, Volume 3, pp. 1486 - 1490, 1999.

[Bec06] Z. Becvar and J. Zelenka, "Handovers in the Mobile WiMAX," Czech Technical University, 2006.

[Don07] G. Dong and J. Dai, "An Improved Handover Algorithm for Scheduling Services in IEEE802.16e," IEEE Mobile WiMAX Symposium, 2007, pp. 38 - 42, March 2007.

[Lee06] D.H. Lee, K. Kyamakya and J.P. Umondi, "Fast handover algorithm for IEEE 802.16e broadband wireless access system," 1st International Symposium on Wireless Pervasive Computing, 2006, January 2006.

[Cho05] S. Choi, G.-H. Hwang, T. Kwon, A.-R. Lim and D.-H. Cho, "Fast handover scheme for real-time downlink services in IEEE 802.16e BWA system," IEEE Vechicular Technology Conference, 2005, Volume 3, pp. 2028 - 2032, May 2005.

[Hu07] R.Q. Hu, D. Paranchych, M-H. Fong, and G. Wu, "On the evolution of handoff management and network architecture in WiMAX," IEEE Mobile WiMAX Symposium, 2007, pp. 144-149, March 2007.

[ASN2] WiMAX Forum Network Architecture (Stage 2: Architecture Tenets, Reference Model and Reference Points - Release 1.1.0).

[ASN3] WiMAX Forum Network Architecture (Stage 3: Detailed Protocols and Procedures - Release 1.1.0).

[Lax06] M. C. Lax and A. Dammander, "WiMAX - A Study of Mobility and a MAC-layer Implementation in GloMoSim," Master's Thesis in Computing Science, Umeå University, April 2006.

[Li06] K-H. Li, "WiMAX Network Architecture," June 2006.

[Wu05] K. Wu, "Load balancing of elastic data streams in cellular networks," Masters's Thesis, Helsinki University of Technology, January 2005.

[Kim07] D. Kim, M. Sawhney, and H. Yoon, "An effective traffic management scheme using adaptive handover time in next-generation cellular networks," International Journal of Network Management, Volume 17, Issue 2, pp. 139 - 154, March 2007.

[Yan05] E. Yanmaz and O.K. Tonguz, "Handover performance of dynamic load balancing schemes in cellular networks," 10th IEEE Symposium on Computers and Communications, 2005, pp. 295 - 300, June 2005.

[Ira00] Y. Iraqi and R. Boutaba, "Resource Management issues in Future Wireless Multimedia Networks," 2000.

[Fuj92] T. Fujii, "Selective handover for traffic balance in mobile communications," ICC'92, Volume 4, pp. 212.3.1–212.3.7, 1992.

[Vel04] H. Velayos, V. Aleo, and G. Karlsson, "Load balancing in overlapping wireless LAN cells," 2004 IEEE International Conference on Communications, Volume 7, pp. 3833 - 3836, June 2004.

[Moi06a] S.N. Moiseev, S.A. Filin, M.S. Kondakov, A. V. Garmonov, A. Y. Savinkov, Y. S. Park, and S. H. Cheon, "Load-Balancing QOS-Guaranteed Handover in the IEEE 802.11 Network," IEEE 17th International Symposium on Personal, Indoor and Mobile Radio Communications, 2006, pp. 1-5, September 2006.

[Moi06b] S.N. Moiseev, S.A. Filin, M.S. Kondakov, A. V. Garmonov, A. Y. Savinkov, Y. S. Park, and S. H. Cheon, "Load-Balancing QoS-Guaranteed Handover in the IEEE 802.16e OFDMA Network," Global Telecommunications Conference, 2006, pp. 1-5, November 2006.

[Lee07] S. H. Lee and Y. Han, "A Novel Inter-FA Handover Scheme for Load Balancing in IEEE 802.16e System," IEEE 65th Vehicular Technology Conference, 2007, pp. 763 - 767, April 2007.

[Ekl86] B. Eklundh, "Channel utilization and blocking probability in a cellular mobile telephone system with directed retry," IEEE Transactions on Communications, Volume 34, Issue 4, April 1986.

[Yum93] T.S. P. Yum and K. L. Yeung, "Blocking and handoff performance analysis of directed retry in cellular mobile systems," IEEE Global Telecommunications Conference, 1993, Volume 1, pp. 537 - 541, November 1993.

[Wat95] F. Watanabe, T. Buott, T. Iwama, and M. Mizuno, "Load sharing sector cells in cellular systems," Sixth IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 1995, Volume 2, pp. 547-551, September 1995.

[Bal02] A. Balachandran, P. Bahl, and G. M. Voelker, "Hot-Spot Congestion Relief in Public-area Wireless Networks," Proc. of 4th IEEE Workshop on Mobile Computing Systems and Applications, June 2002.

[Hong86] H. Daehyoung, and S.S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures," IEEE Transactions on Vehicular Technology, Volume 35, Issue 3, pp. 77 - 92, August 1986.

[Bar04] F. Barcelo, "Performance analysis of handoff resource allocation strategies through the state-dependent rejection scheme," IEEE Transactions on Wireless Communications, Volume 3, Issue 3, pp. 900 - 909, May 2004.

[Lee03] J. Y. Lee, J-G. Choi,K. Park and S. Bahk, "Realistic cell-oriented adaptive admission control for QoS support in wireless multimedia networks," IEEE Transactions on Vehicular Technology, Volume 52, Issue 3, pp. 512 - 524, May 2003.

[Cho98] S. Choi and K. G. Shin, "Predictive and Adaptive Bandwidth Reservation for Hand-Offs in QoS-Sensitive Cellular Networks," ACM SIGCOMM Computer Communication Review, Volume 28, Issue 4, pp. 155 - 166, October 1998.

[Nag95] M. Naghshineh amd M. Schwartz, "Distributed call admission control in mobile or wireless networks," Sixth IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 1995, Volume 1, pp. 289 - 293, September 1995.

[Ira01] Y. Iraqi and R. Boutaba, "When is it worth involving several cells in the call admission control process for multimedia cellular networks," IEEE International Conference on Communications, 2001, Volume 2, pp. 336 - 340, 2001.

[Die04] J. Diederich and M. Zitterbart, "A Simple and Scalable Handoff Prioritization Scheme," Elsevier's Computer Communications, Volume 28, Issue 7, pp. 773-789, May 2005.

[Lev97] D.A. Levine, D.A. Akyildiz, and M. Naghshineh, "A Resource Estimation and Call Admission Algorithm for Wireless Multimedia Networks Using the Shadow Cluster Concept," IEEE/ACM Transactions on Networking, Volume 5, Issue 1, pp. 1 - 12, February 1997.

[Cho00] S. Choi and K. G. Shin, "A comparative study of bandwidth reservation and admission control schemes in QoS-sensitive cellular networks," Wireless Networks, Volume 6, Issue 4, pp. 289-305, 2000.

[Liu98] T. Liu, P. Bahl, and I. Chlamtac, "Mobility Modeling, Location Tracking, and Trajectory Prediction in Wireless ATM Networks," IEEE Journal on Selected Areas in Communications, Volume 16, Issue 6, pp. 922-936, August 1998.

[Jay00] R. Jayaram, S. K. Das, N. K. Kakani, and S. K. Sen, "A call admission and control scheme for quality-of-service (QoS) provisioning in next generation wireless networks," Wireless Networks, Volume 6, Issue 1, pp. 17-30, 2000.

[Zen00] Q.-A. Zeng, D.P. Agrawal, "Performance analysis of a handoff scheme in integrated voice/datawireless networks," 52nd Vehicular Technology Conference, 2000, Volume 4, pp. 1986 - 1992, 2000.

[Xha04] A.E. Xhafa and O.K. Tonguz, "Dynamic priority queueing of handover calls in wireless networks: an analytical framework," IEEE Journal on Selected Areas in Communications, Volume 22, Issue 5, pp. 904 - 916, June 2004.

[Che05] X. Chen, B. Li, and Y. Fang, "A dynamic multiple-threshold bandwidth reservation (DMTBR) scheme for QoS provisioning in multimedia wireless networks," IEEE Transactions on Wireless Communications, Volume 4, Issue 2, pp. 583 - 592, March 2005.

[Chi89] D. Chiu and R. Jain, "Analysis of the Increase and Decrease Algorithms for Congestion Avoidance in Computer Networks," Journal of Computer Networks and ISDN Systems, Volume 17, pp. 1–14, June 1989.

[Jai84] R. Jain, D.M. Chiu and W. Hawe, "A Quantative Measure of fairness and Discrimination for Resource Allocation in Shared Systems," Technical Report, Digital Equipment Corporation, DEC-TR-301, 1984.

[Zha04] H. Zhai, Xiang Chen, Y. Fang, "A call admission and rate control scheme for multimedia support over IEEE 802.11 wireless LANs," First International Conference on Quality of Service in Heterogeneous Wired/Wireless Networks, 2004, pp. 76 - 83, October 2004.

[Bah00] P. Bahl, V. N. Padmanabhan, and A. Balachandran, "A Software System for Locating Mobile Users: Design, Evaluation, and Lessons," February 2000.

[Shr96] M. Shreedhar and G. Varghese, "Efficient fair queuing using deficit round-robin," IEEE/ACM Transactions on Networking, Volume 4, Issue 3, pp. 375 - 385, June 1996.

[Job04] J. Jobin, M. Faloutsos, S.K. Tripathi, S.V. Krishnamurthy, "Understanding the effects of hotspots in wireless cellular networks," 23rd Annual Joint Conference of the IEEE Computer and Communications Societies, 2004, Volume 1, March 2004.

[3GPP2] 3GPP2-TSGC5, HTTP and FTP Traffic Model for 1xEV-DV Simulations.

[Cas07] T. Casey, "Traffic Generation for Evaluation of Mobile WiMAX Scheduler in NS2," Special Assignment in Telecommunications, Communications laboratory, Helsinki University of Technology, 2007.

[Sol06] D. Soldani, M. Li and R. Cuny, "QoS and QoE Management in UMTS Cellular Systems," 2006.

[ns2] http://www.isi.edu/nsnam/ns/

[Ahm06] S. Ahmadi, "Introduction to mobile WiMAX Radio Access Technology: PHY and MAC Architecture," Wireless Standards and Technology, Intel Corporation, December 2006.